Sparse Classifier Fusion for Speaker Verification

Ville Hautamäki, Tomi Kinnunen, Filip Sedlák, Kong Aik Lee, Bin Ma, and Haizhou Li *Senior Member, IEEE*

***Abstract*—State-of-the-art speaker verification systems take advantage of a number of complementary base classifiers by fusing them to arrive at reliable verification decisions. In speaker verification, fusion is typically implemented as a weighted linear combination of the base classifier scores, where the combination weights are estimated using a logistic regression model. An alternative way for fusion is to use classifier ensemble selection, which can be seen as sparse regularization applied to logistic regression. Even though score fusion has been extensively studied in speaker verification, classifier ensemble selection is much less studied. In this study, we extensively study a sparse classifier fusion on a collection of twelve I4U spectral subsystems on the NIST 2008 and 2010 speaker recognition evaluation (SRE) corpora.**

***Index Terms*—Classifier ensemble selection, linear fusion, speaker verification, experimentation**

I. Introduction

SPEAKER verification is the task of accepting or rejecting an identity claim based on a person's speech sample [1]. Modern speaker verification systems utilize ensembles of *base classifiers* to arrive at an accurate verification decision by *classifier fusion*. The base classifiers might utilize, for instance, different speech parameterizations (e.g. spectral, prosodic or high-level features), models (e.g. Gaussian mixture models [2] or support vector machines [3]) or channel compensation techniques (e.g. joint factor analysis [4] or nuisance attribute projection [5]).

In this study, we consider weighted linear combinations of the base classifier scores as the fusion. With a small number of adjustable parameters, linear fusion scheme often shows good generalization performance. But it is crucial for the weights to be optimized using robust method which tolerates reasonable deviations in the base classifier score distributions. In speaker verification, the scores may vary considerably between the training and runtime data mainly due to differences in acoustic environments and transmission channels. The obvious weight optimization strategy, minimizing error rate on the training set, easily overfits [6].

A natural solution is to use a convex surrogate loss function instead that serves as an upper bound to the *0-1 loss function* [6]. Optimizing an upper bound is expected to reduce the classification error rate on the unseen data while strict convexity ensures the existence of a unique global minimum. Well-known loss functions with these desiderata include the *hinge loss* used in training SVMs [7] and the logistic loss, also known as the *logistic regression* model [8]. The latter one has been found reliable in independent studies of fusion in speaker verification [8], [9], [10]. Considered as the *de facto* standard in speaker verification studies, with readily available implementations (e.g. [11], [12]), we take the logistic regression model as our baseline. One further advantage of the model is that the fused scores have an interpretation as automatically calibrated *log-likelihood ratios* (LLRs). In addition to producing interpretable scores, this enables designing the verification threshold using the standard Bayes' minimum risk classifier design [13] based on assumed class priors and pre-specified misclassification costs.

Logistic regression is a probabilistic model of the decision boundary between two classes and its parameters (weights) are usually found as the *maximum likelihood* (ML) estimate on a training set [15]. However, ML solution easily overfits with limited number of training scores (trials) which manifests itself as fusion weights with large magnitude [7]. Consequently, even a small change in the base classifier outputs causes large change in the fusion score leading to unreliable decisions.

Motivated by this observation, we consider *regularized* [16] logistic regression whereby weight vectors with large norm are penalized. Regularization defines a constrained optimization problem where one finds a compromise between training data accuracy while avoiding weights with large magnitude. Regularized solution can also be viewed as *maximum a posteriori* (MAP) estimate of the fusion weights, over which one imposes a prior distribution [16]. As in any practical Bayesian learning method, two additional design concerns are now introduced: (1) choosing the regularizer (functional form of the weight prior) and (2) training its parameters that act as hyperparameters. To exemplify, *ridge regression* [16] or squared Euclidean norm regularization corresponds to choosing an isotropic Gaussian prior with zero mean where the variance parameter determines the degree of regularization applied. In this study, we train the regularization parameters using a held-out validation dataset and focus on the first design question, the choice of the regularizer.

In this paper, we advocate *sparse* regularization applied to logistic regression model training in speaker verification. Sparse regularization means that, in addition to optimizing fusion weights for the full classifier ensemble, we would like to implement simultaneously *classifier ensemble selection* by forcing redundant classifiers to have zero weight. Classifier selection can be also seen as a *feature selection* problem [17]. Feature selection methods are generally divided into three groups: *wrapper* methods that use classification error to select features, *filter* methods that use a surrogate cost function to select features and *embedded* methods which jointly select the subset of features and optimize the classifier parameters.

V. Hautamäki, T. Kinnunen and F. Sedlák are with the Univ. of Eastern Finland, Joensuu, Finland (email: {villeh,tkinnu,fsedlak}@cs.uef.fi). K.A. Lee, B. Ma and H. Li are with Inst. for Infocomm Research (I2R), Singapore (email: {kalee, mabin, hli}@i2r.a-star.edu.sg)

The work of F. Sedlák was supported by Nanyang Technological Univ. (NTU), Singapore, works of T. Kinnunen and V. Hautamäki by Academy of Finland (projects 253000, 253120 and 132129) and the work of H. Li by Nokia foundation.

Manuscript submitted 24.08.2012, revised 21.12.2012

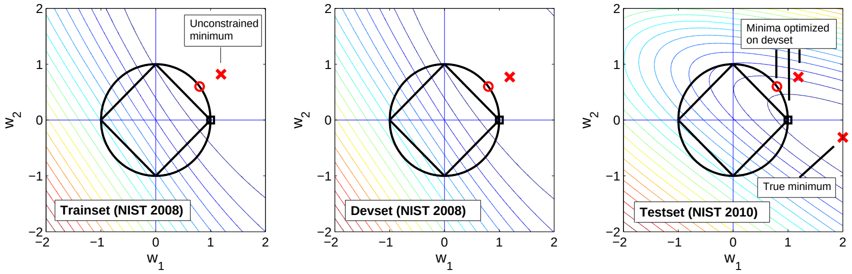


Fig. 1: Regularized classifier fusion. We display the contours of C_{wlr} for fusion of two classifiers. The global minima of C_{wlr} are indicated by red crosses. For constrained optimization, we search for the minimum inside the region $(w_1^p + w_2^p)^{(1/p)} \leq 1$ (here, the cases $p = 1$ and $p = 2$ are displayed). The case $p = 1$ finds a *sparse* solution because classifier 2 gets zeroed. This solution hits closest to the true minimum on the unseen test data. Even ℓ_2 regularization ($p = 2$) outperforms the unconstrained case. Thus, regularization and sparsification might be particularly useful under unpredictable corpus mismatches [14].

Sparse logistic regression studied here belongs to this last category.

There are several arguments favoring the sparse fusion approach. Firstly, even though the full system may consist of up to a dozen of base classifiers (e.g. [18]), these are often redundant; they might utilize only slightly different spectral front-ends, training parameters, acoustic models and development corpora. It is therefore reasonable to assume that the effective number of base classifiers contributing further uncertainty reduction in the fused score is relatively small. An experimental validation for this hypothesis comes from our recent study [19]. Applying exhaustive classifier selection and weight optimization from a pool of 12 classifiers [18], we found that a classifier ensemble with only 4 classifiers outperformed full ensemble in accuracy. Interestingly, similar experimental result was found by MITLL site in their language recognition submission to NIST LRE 2011, subset of 3 base classifiers out of total 5 in the full ensemble gave the best performance [20]. Secondly, reducing the effective number of model parameters is expected to improve generalization performance because of reduced model variance [21]. Finally, computational benefits are obvious during system run-time as the excluded classifiers need not to be invoked.

Even though joint classifier ensemble selection and training the fusion weights is a combinatorial optimization problem, it can be mathematically formulated as ℓ_0-regularization [16] where the regularizer (zeroth norm) counts the number of non-zero weights, corresponding to the selected classifier ensemble. Since its time complexity is still exponential with respect to the number of base classifiers, the usual workaround is to use ℓ_1-regularization instead, a method known as LASSO (*least absolute shrinkage and selection operator*) [22]. In the logistic regression model, ℓ_1-regularization has also been applied [23]. LASSO shrinks all the coefficients, with some of them forced to be exactly zero. By regularizing logistic regression with the LASSO constraint, we can simultaneously optimize fusion weights and perform classifier selection.

Convex combination of ridge regression and LASSO leads to another regularization technique known as *elastic-net* (E-net) [24], which retains the zeroing capability of LASSO, but because of the ridge term it does not push base classifier weights to zero as aggressively as LASSO or classifier ensemble selection do.

This study summarizes and extends our preliminary work on classifier selection [19] and sparse fusion [25], [14]. We expand the theory part in three respects. First, Section II is expanded as a tutorial-like material for readers less familiar with state-of-the-art fusion. Second, we provide mathematical evidence that, under reasonable assumptions, baseline (unregularized) logistic regression is unlikely to produce sparse solutions. Thirdly, we give a detailed account into setting of the regularization parameters. This involves arguing that sparse regularization is able to zero out unreliable classifiers and that, under ideal conditions (no observation noise in the scores), the ℓ_1 solution converges to unregularized solution. The experimentation is further expanded by (1) providing comparison of different score pre-warping variants for calibration purposes, (2) providing detailed comparison of unregularized and regularized fusion schemes on subconditions of the NIST SRE 2010 core task and (3) providing analysis of correlation coefficients across the selected classifiers. To sum up, even though regularization and sparsification have been studied for both linear and logistic regression schemes, their integration into fusion schemes in speaker verification is novel. Our study is the first large-scale comparison of regularized, in particular sparse, fusion schemes in speaker verification.

II. Linear Score Fusion in Speaker Verification

A. Problem Setup

We assume that, during the development phase, one has access to a development set $\mathcal{D} = \{(\mathbf{s}_i, y_i), i = 1, 2, \ldots, N_{\mathrm{dev}}\}$ containing N_{dev} score vectors from L base classifiers, $\mathbf{s}_i \in \mathbb{R}^L$. Here, $y_i \in \{0, 1\}$ indicates whether the corresponding speech sample originates from a target speaker $(y_i = 1)$ or from a non-target $(y_i = 0)$. Though it is not always the case during the NIST SRE campaigns, here we assume that these labels contain no errors. We consider linear score fusion of the form $f_{\mathbf{w}}(\mathbf{s}) = w_0 + \sum_{l=1}^{L} w_l s_l = \mathbf{w}^{\mathsf{T}}\mathbf{s}$, where $\mathbf{w} = (w_0, w_1, \ldots, w_L)^{\mathsf{T}}$ contains the classifier weights $w_1, \ldots, w_L$ (discrimination component) and the bias w_0 (calibration component). The augmented score vector $\mathbf{s} = (1, s_1, s_2, \ldots, s_L)^{\mathsf{T}}$ contains constant 1 and the base classifier output scores.

Our goal is to find optimal weight vector (say, $\mathbf{w}^*$) so that classification errors are minimized on the development data, thus hopefully on the unseen evaluation data. Here we adopt the *detection cost function* (DCF) commonly used in the NIST speaker recognition evaluations[1] to assess the accuracy of any speaker verification system:

$$\text{DCF}(\theta) = C_{\text{miss}} P_{\text{miss}}(\theta) P_{\text{tar}} + C_{\text{fa}} P_{\text{fa}}(\theta)(1 - P_{\text{tar}}). \quad (1)$$

Here, $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$ are the miss and false alarm probabilities as a function of the decision threshold θ, P_{tar} is the prior probability of a target (true) speaker, C_{miss} is the cost of a miss (false rejection) and C_{fa} is the cost of a false alarm (false acceptance).

In speaker verification, (1) is used for computing both the *actual* (ActDCF) and *minimum* (MinDCF) values. The actual cost refers to the DCF value obtained whenever the decision threshold θ is fixed to a particular value beforehand, whereas MinDCF indicates the oracle value (minimum) on the test set that can easily be found by linear search over the range of θ. Therefore, by definition ActDCF $\geq$ MinDCF, and the difference ActDCF - MinDCF can be used as a measure of calibration error in terms of how well the w_0 was estimated. Score magnitudes can also affect calibration, as is the case in NIST SRE 2012, but in this present work we consider only the additive term.

B. Logistic regression

To train the fusion device, in theory one can optimize (1) directly, for instance by using a neural network [26]. For the reasons discussed above, we optimize the weights using a convex loss function instead. *Logistic regression* is a probabilistic linear model, which is based on the fact that posterior probability of the class label being the target class can be written as $p(y = 1|\mathbf{s}) = (1 + \exp\{-g(\mathbf{s})\})^{-1}$ for *any* class-conditional densities [15]. The function $g(\mathbf{s})$ takes the form $\mathbf{w}^{\mathsf{T}}\mathbf{s}$ when the class-conditional densities follow exponential family of distributions with a shared dispersion parameter (e.g. variance). We can thus express the target class posterior as $p(y = 1|\mathbf{s}) = (1 + \exp\{-(\mathbf{w}^{\mathsf{T}}\mathbf{s})\})^{-1} = \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{s})$ [15], where $\sigma(.)$ is a logistic sigmoid function. The posterior for the non-target class is then $p(y = 0|\mathbf{s}) = 1 - p(y = 1|\mathbf{s}) = \sigma(-\mathbf{w}^{\mathsf{T}}\mathbf{s})$ by the properties of $\sigma(.)$. Furthermore, the quantity $\mathbf{w}^{\mathsf{T}}\mathbf{s}$ has an interpretation as the *log odds,* i.e. $\ln[p(y = 1|\mathbf{s})/p(y = 0|\mathbf{s})]$ [7], as one can verify by straightforward algebra.

Using the development above, we are now able to write the likelihood function for the logistic regression model [7]:

$$p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^{N_{\text{dev}}} \left\{ \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{s}_n)^{y_n} \sigma(-\mathbf{w}^{\mathsf{T}}\mathbf{s}_n)^{1-y_n} \right\}, \quad (2)$$

where $\mathbf{y}$ is the N_{dev}-dimensional vector of all labels y_n. Maximum likelihood (ML) estimate of $\mathbf{w}$ can be found by taking the negative logarithm of (2), which yields the *cross-entropy* cost [7]:

$$-\sum_{n=1}^{N_{\text{dev}}} \left\{ y_n \ln \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{s}_n) + (1 - y_n) \ln \sigma(-\mathbf{w}^{\mathsf{T}}\mathbf{s}_n) \right\}. \quad (3)$$

This is also known as the C_{llr} cost in [27]. The minimum of (3) does not have closed form soluton [7], however it is convex, so iterative gradient descent methods can be used to find optimal $\mathbf{w}^*$.

The above formulation assumes that the costs of miss and false alarm are equal ($C_{\mathrm{miss}} = C_{\mathrm{fa}}$) and that $P_{\mathrm{tar}} = 0.5$. To re-calibrate the model according to the pre-specified cost parameters (C_{miss}, C_{fa} and P_{tar}), the following modification is used [27]:

$$p(y = 1|\mathbf{s}) = \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{s} + \operatorname{logit} P_{\mathrm{eff}}), \quad (4)$$

where P_{eff} is known as *effective prior,* which summarizes the three application-dependent parameters into a single parameter, $P_{\mathrm{eff}} = \operatorname{logit}^{-1}(\operatorname{logit}(P_{\text{tar}}) + \log(C_{\text{miss}}/C_{\text{fa}}))$, with $\operatorname{logit} P = \log(P/(1 - P)) = -\theta$. Bayes-optimal decision is then achieved by placing the threshold to $-\operatorname{logit}(P_{\mathrm{eff}})$.

In addition to DCF parameters, the number of positive and negative examples in the development set might be highly imbalanced. This is the case with the NIST evaluations. As an example, in the female itv-itv condition in NIST SRE 2010 only 3.45% of the trials are target (positive) trials. This would mean that the cross-entropy objective (3) will be strongly dominated by one of the two classes leading to biased weights. To take this class imbalance problem into account, the cost was further modified in [9] as follows:

$$\begin{aligned} C_{\mathrm{wlr}}(\mathbf{w}, \mathcal{D}) &= \frac{P_{\mathrm{eff}}}{N_t} \sum_{i=1}^{N_t} \log\left(1 + e^{-\mathbf{w}^{\mathsf{T}}\mathbf{s}_i - \operatorname{logit} P_{\mathrm{eff}}}\right) \\ &+ \frac{1 - P_{\mathrm{eff}}}{N_f} \sum_{j=1}^{N_f} \log\left(1 + e^{\mathbf{w}^{\mathsf{T}}\mathbf{s}_j + \operatorname{logit} P_{\mathrm{eff}}}\right) (5) \end{aligned}$$

where the two sums go through the N_t target score vectors $\mathbf{s}_i$ and the N_f non-target score vectors $\mathbf{s}_j$, respectively.

C. Score pre-warping

Since the raw base classifier scores may have different interpretations (e.g. log-likelihood ratios, SVM scores or i-vector cosine distances) with considerable variation in their scales, it is important to properly align the score distributions [28]. Note that the base classifiers typically include their internal score normalization such as T-norm [29], used for normalizing the classifier outputs across varying test segments and speakers with the help of external cohort models. Here the concern is to make global score alignment at the classifier level. To avoid confusion with speaker score normalization techniques, we refer to global classifier-level score pre-processing as *score pre-warping.*

Most common pre-warping is *mean and variance normalization* (MVN), also known as z-normalization. Mean (μ) and standard deviation (σ) of the entire score distribution is estimated from the training data and applied to the held out score (s) as $s \mapsto (s - \mu)/\sigma$. MVN defines affine score normalization whose parameters can also be discriminatively learned, as we will see later.

In addition to the MVN where the range of the pre-warping function was unbounded, we also consider *z-cal* [30] and *s-cal* [9] methods that intentionally set upper and lower limits

[1] http://www.itl.nist.gov/iad/mig/tests/spk/

TABLE I: Score pre-warping methods used in this study.

Type of pre-warping	# parameters	Discr. Learning
MVN	2	No
z-cal (unclip)	2	Yes
z-cal (clip)	4	Yes
s-cal	4	Yes

on the pre-warped scores. We thus call these methods *clipped* variants. These methods were originally devised to overcome the problem of labeling errors, assumption being that small portion of target trials were accidentally marked as non-target and vice versa. In [9], s-cal was applied on the fusion *output* but we apply it to the *inputs* before fusion. By applying clipping to the score pre-warping, non-linearity is applied. This leads to a score pre-warping effect that linear fusion device is not able to recreate.

Both z-cal and s-cal aim at converting arbitrary scores to well-calibrated log-likelihood ratios (LLRs). The s-cal pre-warping is

$$\mathrm{LLR}_{\mathrm{scal}}(s_n) = \log \frac{(\mathrm{logit}^{-1}\alpha)(e^{xs_n+y}-1)+1}{(\mathrm{logit}^{-1}\beta)(e^{xs_n+y}-1)+1}, \quad (6)$$

where the saturation parameters α, β and the affine parameters x, y are optimized using the development set, with the attached ground truth labels so that the C_{llr} cost in (3) is optimized [27]. As the problem is no longer convex in the unknowns, we utilize unconstrained nonlinear Nelder-Mead optimization algorithm [31] to find locally optimum values for α, β, x and y. In each new estimate of the parameters, the development set scores are pre-warped using (6) and the optimality of the parameter estimates is computed using C_{llr} in (3); we utilize Matlab's `fminsearch` function to implement this. Occassionally the optimizer produced singular solutions. Those were detected, by noticing that then C_{llr} is one, and rejected. If a solution was rejected then new one is computed by stronger regularization.

The z-cal pre-warping function is defined similarly to s-cal, only difference being that instead of smooth sigmoidal shape, z-cal defines a piece-wise linear function with hard thresholding (clipping). Z-cal is defined as:

$$\mathrm{LLR}_{\mathrm{zcal}}(s_n) = (s_n - x_{\min})\frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}} + y_{\min}, \quad (7)$$

where we set $\mathrm{LLR}_{\mathrm{zcal}}(s) = y_{\min}$ for all scores satisfying $\mathrm{LLR}_{\mathrm{zcal}}(s_n) < y_{\min}$; similarly $\mathrm{LLR}_{\mathrm{zcal}}(s_n) = y_{\max}$ for all scores with $\mathrm{LLR}_{\mathrm{zcal}}(s_n) > y_{\max}$. z-cal parameters are optimized in a same way as s-cal parameters. We also experimented with the unclipped variant of z-cal, optimization was performed in a same way except that the clipping step was not used. It is expected that unclipped z-cal will provided similar results as optimizing fusion scores without pre-warping. Score pre-warping methods selected for this study have been summarized in Table I.

III. Regularized Classifier Fusion

A. Unregularized model is unlikely to find sparse solutions

We argue that in order to produce, in a general case, a sparse weight vector, $\mathbf{w}$ one has to use sparsity promoting regularizer, such as ℓ_0 and ℓ_1. In the following, we see when unregularized logistic regression (i.e. via maximum likelihood training) can produce sparse solutions. First, the maximum likelihood solution of $\mathbf{w}$ is characterized as [15]:

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \quad (8)$$

if the class-conditional densities follow Gaussian distribution. In (8), $\mathbf{\Sigma}$ is the shared covariance matrix and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ are the class-conditional mean vectors. If we take $\mathbf{\Sigma}$ to be diagonal, as was assumed in [8], then for each base classifier,

$$w_l = \frac{1}{\sigma_l^2} d_l, \quad l = 1, \ldots, L \quad (9)$$

where d_l is the difference between the means for the lth dimension. It is clear that, under these assumptions, w_l can be exactly zero only when the means of target and non-target scores completely match which is unlikely to happen for any reasonably-performing speaker verification system. On the other hand, an extremely large variance σ_l^2 would also push w_l arbitrarily close to zero, but not exactly zero.

The above argument assumed diagonal covariance matrices and a particular special case of logistic regression. Even though the same analysis no longer holds for full covariance matrices, it does illustrate that there are cases when unregularized solution cannot find a sparse solution. With sparsity promoting regularizers, on the other hand, we can force sparse weights regardless of whether the classifiers are correlated or not.

B. Classifier Ensemble Selection as Regularization

Up to this point, we have defined the standard fusion framework, assuming a full ensemble of L classifiers. Now, instead of just optimizing the weights, we are in search of both the optimal classifier ensemble and weights. Assume that we have decided on an appropriate size of the ensemble given by integer variable K, $K < L$. We would like to minimize (5) subject to this constraint. Obviously, one can simply enumerate all the $\binom{L}{K} = \frac{L!}{K!(L-K)!}$ possible classifier ensembles to ensure that the size constraint is satisfied, and optimize the weights by minimizing C_{wlr} for each of these ensemble candidates and choosing the one that minimizes the cost function.

In this paper, we show that a better way formulating the problem is by casting it into the regularization framework. Interestingly, the exhaustive search can be seen as regularization with ℓ_0-norm. Note that the ℓ_0-norm of vector $\mathbf{w}$, defined as $\|\mathbf{w}\|_0 \triangleq \sum_i |w_i|^0$, counts the number of nonzero elements in $\mathbf{w}$ [16]. This is because $|w_i|^0$ equals 1 everywhere except for $w_i = 0$ one defines it as 0. Thus, an equivalent formulation of the above combinatorial optimization problem is

$$\min_{\mathbf{w}} \; C_{\mathrm{wlr}}(\mathbf{w}, \mathcal{D}) \quad \text{s.t.} \quad \|\mathbf{w}\|_0 \leq K. \quad (10)$$

Although it is clear that the combinatorial search outlined above is not very practical for large L, it is guaranteed to give the optimum classifier ensemble choice for a given set of data.

C. Practical Regularization via Ridge, LASSO and Elastic Net

For computational reasons the $\|\mathbf{w}\|_0$ constraint is typically approximated using the $\|\mathbf{w}\|_1$ constraint, which is also known as LASSO [16]. The vector norm can also be constrained by $\|\mathbf{w}\|_2^2$, which corresponds to *ridge regression*. However, unlike LASSO, ridge is not a sparsity promoting constraint [16].

In the case of LASSO, (10) is modified as,

$$\min_{\mathbf{w}} \; C_{\mathrm{wlr}}(\mathbf{w}, \mathcal{D}) \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq t, \tag{11}$$

where t determines the desired amount of shrinkage. In (10), the norm $\|\mathbf{w}\|_0 \in \mathbb{Z}^+$ has an interpretation as the maximum number of classifiers retained, but this does not necessarily hold for $\|\mathbf{w}\|_1 \in \mathbb{R}^+$ which takes up any positive real value. Therefore, rather than based on human judgment, t in (11) should be merely considered as a control parameter. In this study, we optimize t using cross-validation. A useful insight into choosing a suitable range of possible t's is the desired amount of shrinkage *relative* to the unregularized solution. That is, set $t = \hat{t} \cdot \|\mathbf{w}^{\mathrm{ML}}\|_1$, where $\mathbf{w}^{\mathrm{ML}}$ are the maximum likelihood weights for (5) and $\hat{t}$ is the desired amount of shrinkage, such as $\hat{t} = 0.90$.

From a viewpoint of optimization software packages, a more useful form of (11) is its Lagrange multiplier formulation,

$$\min_{\mathbf{w}} \; \{C_{\mathrm{wlr}}(\mathbf{w}, \mathcal{D}) + \lambda \|\mathbf{w}\|_1\}, \tag{12}$$

where λ is the Lagrange multiplier. It is known that the larger λ, the more the norm $\|\mathbf{w}\|$ will be shrunk [22]. Example of (12) on real data can be seen in Fig. 1, where two base classifiers are fused. From the example it is clear that weights found by the direct optimization of (5) would lead to non-optimal solution for the test set.

When optimization is based on (12), the correspondence between λ and the shrinkage threshold t can be found by a binary search on the possible values of λ. In each iteration, we select one λ and optimize the weights using it, output is then the norm of the weights. Final weight vector is the one whose norm is closest to the target t, but does not violate it.

Elastic-net, on the other hand, is based on the idea that we can combine both ℓ_1 and ℓ_2 regularizers into one constrained optimization problem,

$$\min_{\mathbf{w}} \left\{ C_{\mathrm{wlr}}(\mathbf{w}, \mathcal{D}) + \lambda \left(\alpha \|\mathbf{w}\|_1 + (1 - \alpha)\|\mathbf{w}\|_2^2 \right) \right\}. \tag{13}$$

As can be seen, Eq. (13) is a generalized variant of both LASSO and ridge regression. One can always find such a α where, in terms of performance, elastic-net will at least as good as LASSO or ridge regression. However, whereas LASSO and ridge regression had to select only one regression parameter, now we need to cross-validate over a 2-d space. In this work, the α parameter is first fixed and then shrinkage factor λ is cross-validated as in LASSO and ridge. In practice, α will also be cross-validated in so that the best α and shrinkage factor will be selected based on cross-validation set to be applied on the evaluation set.

TABLE II: Selection of the three datasets used in this study.

Dataset	Usage	Data source	# Trials
Trainset	Train fusion parameters	NIST 2008 itv-itv ♀subset	2434 t, 238971 f
Devset	Compare fusion and pre-warping methods and classif. selection	NIST 2008 itv-itv ♀subset	2408 t, 239244 f
Evalset	Validate results	NIST 2010 itv-itv ♀subset	5235 t, 146623 f

Depending on the chosen regularization method, there are different strategies to optimize regularized cross-entropy objective. Since logistic regression using quadratic regularization is differentiable, it can be efficiently optimized using standard packages [7]. Situation is not so simple for LASSO regularization. In [22], a *quadratic programming* (QP) solution was proposed by rewriting the constraints in (12) to a more convenient form. However, more recent techniques are faster in practice, for that reason we apply *projectionL1* algorithm [32][2] that optimizes the Lagrangian form of (12). We apply the same method to elastic-net. Since the sum of two convex functions is still convex, we can minimize $C_{\mathrm{wlr}}(\mathbf{w}, \mathcal{D}) + \lambda(1-\alpha)\|\mathbf{w}\|_2^2$, given $\lambda\alpha\|\mathbf{w}\|_1$ as the constraint.

D. Sparse Regularization Knocks Out Noisy Classifiers

Is there a way to tell whether a given base classifier gets knocked out for a given λ? In the case of ℓ_1-norm and standard linear regression, it has been shown that the weights satisfy [33],

$$w_l = \begin{cases} w_l^{\mathrm{ML}} - \epsilon_l \, \mathrm{sign}(w_l^{\mathrm{ML}}), & |w_l^{\mathrm{ML}}| > \epsilon_l \\ 0, & |w_l^{\mathrm{ML}}| \leq \epsilon_l \end{cases}, \tag{14}$$

where w_t^{ML} is the maximum likelihood estimate of the weight of the base classifier l, $\epsilon_l = \frac{\sqrt{\lambda}\sigma^2}{\sum_n s_{ln}^2}$, σ^2 is noise variance (needed in standard regression setup) and s_{ln} is the nth score of base classifier l. Thus, ℓ_1-norm based regularization defines an interval of $2\epsilon_l$ around the origin where zero weight is obtained. For the case of logistic regression, ℓ_1 regularization similarly defines an interval around the origin, where weight is zero [33]. Note that we can write $\epsilon_l = \sqrt{\lambda}/\xi_l$, where $\xi_l \triangleq \sum_n s_{ln}^2/\sigma^2$ is the ratio of "signal" to noise variance. Therefore, the noisier the classifier scores (lower ξ_l), the larger the interval $2\epsilon_l$ and the higher the chance that a classifier gets zeroed out. For noise-free case ($\xi_l \to \infty$), the solution converges to the ML weight: $w_l \to w_l^{\mathrm{ML}}$.

This above reasoning shows that, while constant λ is applied for all base classifiers, the amount of shrinkage (hence, zeroing out) depends on the noise level of that base classifier. Since the noise variance is generally unknown, λ can be used for adjusting the zeroing interval with cross-validation. The reasoning here concerns standard linear and logistic regression but similar arguments could be made for regularized C_{wlr}.

IV. Corpora, Metrics and Base Classifiers

We utilize the NIST 2008 and NIST 2010 corpora in our experiments[3]. The usage of each corpus is shown in Table II.

[2] http://www.di.ens.fr/~mschmidt/Software/code.html

[3] http://www.itl.nist.gov/iad/mig/tests/sre/

TABLE III: Twelve base classifiers, calibrated using MVN, are constructed using different cepstral features and speaker modeling techniques.

	Classifier	Feature	Devset				Evalset			
			EER (%)	MinDCF ($\times$100)	ActDCF ($\times$100)	ActDCF-MinDCF	EER (%)	MinDCF ($\times$100)	ActDCF ($\times$100)	ActDCF-MinDCF
1	GMM-UBM-JFA	PLP	3.44	1.6748	1.6979	0.0231	7.18	3.3108	3.3911	0.0803
2	GMM-UBM-JFA	PLP	3.45	1.4309	1.5547	0.1238	5.74	2.3852	2.4268	0.0416
3	GMM-UBM-JFA	PLP	3.32	1.4760	7.7305	6.2545	4.62	2.6668	8.2292	5.5624
4	GMM-UBM-JFA	LPCC	3.99	1.9056	7.8119	5.9063	10.68	5.7845	6.5031	0.7186
5	GMM-SVM-KL	PLP	3.74	1.8597	5.2105	3.3508	6.82	2.9659	6.9683	4.0023
6	GMM-SVM-KL	MFCC	3.16	1.1564	1.4921	0.3357	5.45	2.7169	2.7338	0.0168
7	GMM-SVM-KL	LPCC	3.53	1.4877	1.8412	0.3535	8.35	4.1369	6.2928	2.1559
8	GMM-SVM-KL	MLF [34]	2.95	1.2965	1.7472	0.4508	8.29	3.9229	4.4433	0.5204
9	GMM-SVM-KL	LPCC	3.82	1.9267	5.2591	3.3324	10.55	4.9308	4.9947	0.0639
10	GMM-SVM-KL	SWLP [35]	6.69	3.6348	3.6585	0.0237	10.75	5.0897	5.7239	0.6342
11	GMM-SVM-FT [36]	PLP	4.45	1.9574	6.6046	4.6472	8.60	3.7126	8.0517	4.3391
12	GMM-SVM-BHAT [37]	PLP	3.12	1.2151	1.3090	0.0938	6.28	2.9944	3.0175	0.0232

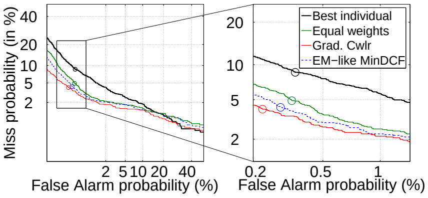


Fig. 2: Comparison of fusion methods using the full ensemble s-cal pre-warping on Trainset. The best individual classifier (for ActDCF) is also shown. The circles indicate the ActDCF points.

To avoid any evaluation bias from pooling of incompatible subcondition scores (see [38]), we mostly focus on the female trials[4] of the interview-interview (itv-itv) sub-condition in the core task. Nevertheless, both genders and three other sub-conditions (itv-tel, mic-mic, tel-tel) are included into the final validation. The audio files from all NIST 2008 speakers were split into two disjoint parts. In this regard, audio files (including both training and test files in the official NIST 2008 SRE dataset) from the same speaker were grouped together based on the available metadata. We then splitted the speakers into two groups, consisting of 475 and 711 speakers, respectively. Trials were then generated separately from those two sets by assigning training and test files randomly based on the speaker information. We kept the empirical $P_{\rm tar}$ close to those in the official NIST 2008 SRE trial lists. The first part, *Trainset*, is used for training the score pre-warping parameters (s-cal was used as precalibration method), fusion weights and bias. The second part, *Devset*, is used for optimizing the ensemble size (K) for subset selection, shrinkage parameter (λ) for LASSO, ridge and elastic net, and the tradeoff between LASSO and ridge for elastic net (α). The optimized parameters are then applied to the NIST SRE 2010 trials (*Evalset*), which serves for evaluation purposes. For the oracle subset selection, the classifier ensembles are optimized by exhaustive search on Evalset.

For evaluation of the methods, we consider the detection cost function in (1), where the cost parameters are $C_{\rm miss} = 10$, $C_{\rm fa} = 1$ and $P_{\rm tar} = 0.01$. We measure both the minimum DCF (MinDCF) and the actual DCF (ActDCF). We also consider *calibration error*, defined as the difference of ActDCF and MinDCF, and the well-known *equal error rate* (EER), corresponding to the case of equal miss and false alarm rates[5].

Table III shows our twelve base classifiers based on different cepstral features and four different speaker modeling techniques. When a base classifier shares the same model and features, it means that the base classifiers are independent implementations. For speaker modeling, we use the generative GMM-UBM-JFA [4] and the discriminative GMM-SVM approaches with KL-divergence kernel [39] and Bhattacharyya kernel (BHAT) [37]. We also include feature transformation (FT) method [36] as an alternative supervector for SVM. All of the methods are grounded on the universal background model (UBM) paradigm [2] and share similar form of subspace channel compensation, though the training methods differ. We used data from the NIST 2004, 2005 and 2006 corpora to train the UBM and the session-variability subspaces, and additional data from the Switchboard corpus to train the speaker-variability subspace for the JFA systems. Each base classifier has its own score normalization prior to score pre-warping and fusion. To this end, we use TZ-norm [29] with NIST 2004 and NIST 2005 data as the background and cohort training data.

V. Results

A. Choosing Score Pre-Warping and Fusion Training Methods

We first compare the score pre-warping and fusion training methods on the full set of $L = 12$ base classifiers in Table IV. Here we consider three methods to obtain the fusion method. The first method, *equal weights* uses uniform weights and does not require training. In the second method, *Gradient* $C_{\rm wlr}$, we use standard conjugate gradient optimization of the weighted

[4] Female trials are somewhat more difficult than males. Similar rationale was taken, for instance, in [4].

[5] For finite data points, one does not find $P_{\rm miss} = P_{\rm fa}$ exactly. In this study, we use linear interpolation between the two closest discrete data points to compute EER. For the interested reader we point to the alternative method using convex hulls on ROC curve (ROCCH), available in [12].

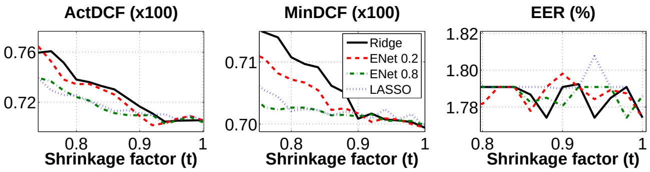
logistic regression cost, $C_{\rm wlr}$. In the third method, *EM-like MinDCF*, we directly optimize MinDCF using an EM-like procedure as follows. We start with equal weights and find the threshold θ that minimizes MinDCF. Given fixed threshold, weights are optimized using the Nelder-Mead algorithm [31]. The process is iterated until convergence; for more details, refer to [19].

The first three rows show best individual base classifiers in terms of ActDCF, MinDCF and EER. As these scores are not pre-calibrated, calibration error is quite large. As expected, fusion improves accuracy over the best single classifier systematically. Regarding score pre-warping, z-cal and s-cal yield similar results. They produce less errors compared to both the unwarped and the non-clipped score pre-warping variants. Fusion training with Grad. $C_{\rm wlr}$ and with no score pre-warping at all and unclipped z-cal yields same EER and ActDCF, but in MinDCF there is a slight difference. As the optimization cost of linear calibration and $C_{\rm wlr}$ are slightly different, there are small differences in MinDCF. In addition, generative pre-warping strategy by MVN also yields different but comparable results to all three unclipped variants.

Comparing the fusion training methods, gradient $C_{\rm wlr}$ systematically outperforms the other two methods in all three costs. The DET plot in Fig. 2 confirms this. We find the direct optimization of MinDCF produces generally higher error rates than logistic regression ($C_{\rm wlr}$) which does only indirect minimization. This suggests that logistic regression offers better generalization performance. For the rest of the experiments, we choose gradient $C_{\rm wlr}$ with s-cal.

TABLE IV: Fusion of all the $L = 12$ base classifiers on the Devset. The first three rows show the individually best base classifiers.

Fusion method	Score pre-warping	EER	MinDCF	ActDCF	ActDCF-MinDCF
Best ActDCF	–	3.74	1.8597	**3.0131**	**1.1534**
Best MinDCF	–	3.16	**1.1564**	18.4600	16.600
Best EER	–	**2.95**	1.2965	14.7607	13.464
Equal weights	–	2.09	0.8385	5.9863	5.1478
	MVN	2.10	0.8219	2.3085	1.4865
	z-cal (unclip)	2.08	0.8080	1.1022	0.2942
	s-cal	**2.03**	0.7907	**0.9176**	**0.1269**
	z-cal (clip)	1.99	**0.7786**	0.9617	0.1830
Grad. $C_{\rm wlr}$	–	1.83	0.6172	0.6231	**0.0059**
	MVN	1.83	0.6139	0.6235	0.0096
	z-cal (unclip)	1.83	0.6135	0.6231	0.0096
	s-cal	1.70	0.6031	**0.6147**	0.0116
	z-cal (clip)	**1.66**	**0.5940**	0.6183	0.0243
EM-like MinDCF [19]	–	2.03	0.7038	2.1931	1.4892
	MVN	2.03	0.7095	4.2973	3.5878
	z-cal (unclip)	2.03	0.7159	**1.5044**	**0.7885**
	s-cal	**1.89**	**0.6440**	2.7454	2.1014
	z-cal (clip)	1.95	0.6631	9.9502	9.2871



(a) (b) (c)

Fig. 3: Effect of the shrinkage factor (Devset, female trials, itv-itv condition). The relative shrinkage factor in x-axis is $\hat{t}$ in Eq. (11) normalized by the unregularized weight norm.

TABLE V: Chosen NIST 2010 subconditions.

	NIST 2010 common cond.								
	1	2	3	4	5	6	7	8	9
itv-itv	×	×							
itv-tel			×						
mic-mic							×		×
tel-tel					×	×		×	

B. Effect of Regularization

We now turn attention to weight optimization using the three regularizers described above (ridge, LASSO and elastic net). Fig. 3 shows the effect of regularization to recognition accuracy on Devset. For ease of interpretation, we show the accuracy as a function of the normalized regularization contraint $\hat{t} = t/\|\mathbf{w}^{\rm ML}\|$ rather than the Lagrange multiplier λ. Here, t is the constraint in Eq. (11) and $\mathbf{w}^{\rm ML}$ denotes the unregularized (maximum likelihood) weight vector. Thus, $\hat{t} = 1$ corresponds to the unregularized solution.

In sparse regularized fusion training, all weights are constrained by regularizer, some are pushed to zero, but even those that are retained are regularized. Thus, when mismatch between Devset and Trainset is small it is expected that even subset of classifiers, which weights are regularized, cannot improve on the unregularized fusion. Elastic net with $\alpha = 0.2$ marginally improves ActDCF. Elastic net with $\alpha = 0.2$ and $\alpha = 0.8$, respectively, has similar ActDCF trends as ridge and LASSO, as one may expect. A general trend is that aggressive shrinking (small $\hat{t}$) increases both MinDCF and ActDCF. The equal error rate (EER), however, does not follow the same trend; this might be because weight optimization target is the DCF rather than the EER region.

C. Extended Results on Other Conditions

Table VI shows accuracies for all the subconditions of the NIST 2010 core task as listed in Table V. We compare five fusion strategies:

- **Best individual**: individually best base classifier (smallest ActDCF on Devset)
- **No regularization**: unregularized logistic regression, similar to FoCal and BOSARIS software packages.
- **Ridge**: ridge regression (ℓ_2) regularization.
- **LASSO**: LASSO (ℓ_1) regularization.
- **E-net**: elastic net (ℓ_2 and ℓ_1) regularization.

All of these are treated the same way regarding the use of datasets: fusion training is carried out on Trainset while the regularization parameters are optimized on Devset, where minimum ActDCF is used as the criterion. In the case of ties, we select the most aggressive regularization factor. Optimizations

TABLE VI: Full comparison of fusion methods on NIST SRE 2010. All fusion parameters have been cross-validated using Devset. The star ($\star$) denotes a statistically significant difference (McNemar's test [40], [41] at 95 % confidence) to unregularized fusion regarding the number of misses (N_{miss}) or false alarms (N_{fa}). The $\hat{t}$ is the normalized shrinkage constraint relative to unregularized norm. The total number of genuine (N_{gen}) and impostor (N_{imp}) trials in each condition are also indicated.

	Ensemble selection	EER (%)	(N_{miss},N_{fa})	MinDCF (×100)	(N_{miss},N_{fa})	ActDCF (×100)	(N_{miss},N_{fa})	ActDCF-MinDCF	$\hat{t}$	Ensemble Size	$N_{\text{gen.}}$	$N_{\text{imp.}}$
itv-itv	Best ind.	5.45	(285, 7997)	2.7169	(972, 1274)	3.8767	(436, 4508)	1.1597	1	1	5235	146623
	No regul.	3.41	(179, 4996)	1.7135	(581, 894)	2.5198	(270, 2968)	0.8063	1	12		
	Ridge	3.40	(178, 4986)	1.7012	(594, $\star$839)	2.5109	(271, 2952)	0.8097	0.96	12		
	LASSO	**3.32**	(174, $\star$4865)	**1.6869**	(595, $\star$815)	**2.2354**	(288, $\star$2496)	**0.5486**	0.96	6		
	E-net $\alpha = 0$	3.40	(178, 4986)	1.7012	(594, $\star$839)	2.5109	(271, 2952)	0.8097	0.96	12		
itv-tel	Best ind.	3.03	(24, 918)	1.3879	(75, 138)	1.7761	(50, 352)	0.3882	1	1	801	30254
	No regul.	2.45	(19, 742)	0.9773	(56, 85)	1.7102	(29, 412)	0.7330	1	12		
	Ridge	2.40	(19, 726)	**0.9689**	(59, $\star$71)	1.6513	(29, $\star$394)	0.6824	0.86	12		
	LASSO	2.37	(18, $\star$716)	0.9865	(57, 84)	1.6332	(32, $\star$377)	0.6467	0.71	8		
	E-net $\alpha = 0.7$	2.37	(19, $\star$718)	0.9746	(55, 88)	**1.4740**	(30, $\star$336)	0.4994	0.66	10		
mic-mic	Best ind.	6.52	(23, 2068)	3.0379	(61, 420)	3.1569	(75, 331)	0.1190	1	1	353	31744
	No regul.	5.12	(18, 1625)	2.3549	(61, 201)	4.4200	(27, 1172)	2.0651	1	12		
	Ridge	5.10	(18, 1618)	**2.2964**	(64, $\star$155)	3.0418	(40, $\star$612)	0.7454	0.66	12		
	LASSO	5.62	(20, 1785)	2.4412	(57, 265)	3.2276	(35, $\star$717)	0.7864	0.56	3		
	E-net $\alpha = 0.7$	**4.82**	(17, $\star$1529)	2.3086	(63, $\star$168)	**3.0330**	(42, $\star$591)	0.7243	0.51	6		
tel-tel	Best ind.	3.62	(26, 1763)	1.5782	(85, 195)	1.6563	(82, 254)	0.0781	1	1	719	48753
	No regul.	**2.36**	(17, 1153)	**1.1151**	(52, 193)	**1.1980**	(60, 179)	0.0828	1	12		
	Ridge	**2.36**	(17, 1153)	1.1422	(50, 220)	1.2133	(63, $\star$166)	0.0712	0.91	12		
	LASSO	**2.36**	(17, 1153)	1.1810	(49, 246)	1.2761	(70, $\star$149)	0.0951	0.91	5		
	E-net $\alpha = 0.1$	**2.36**	(17, 1153)	1.1364	(49, 224)	1.2153	(63, $\star$167)	0.0790	0.81	12		

are carried out separately for each of the four subconditions using their Trainset and Devset counterparts.

We make several interesting observations from Table VI. Firstly, comparing the best individual classifier to the other strategies, fusion of multiple base classifiers outperforms individual classifier in nearly all the cases. In a few cases (most notable, itv-tel), the single classifier has good calibration though. Second, comparing the unregularized baseline to the regularized variants, one of the latter variants wins in most conditions. The exception is the tel-tel condition where the unregularized baseline outperforms all the regularized variants. In fact, tel-tel condition is the easiest case, possibly due to larger development set and longer experience of the team in processing telephony data.

Comparing ridge, LASSO and elastic net, none is a clear winner but the relative performance depends on the condition and metric. Regarding the primary metric, ActDCF, all of them are useful for reducing the number of false alarms compared to the unregularized baseline by a statistically significant margin. For instance, with only a slight increase of target speaker misses, ridge and elastic net reduce the number of false alarms to nearly half of that of the unregularized baseline on the mic-mic condition. Generalization bounds show that sparse solutions that give low error rates have a good chance of generalizing to an unseen dataset [42]. However, as such bounds are loose on non-sparse solutions, depending on the data set, dense weight weight vector can generalize well also as we have seen here.

TABLE VII: Pearson's correlation analysis of sparse fusion ensemble on elastic net method for the mic-mic condition. Correlation is computed between base classifier scores retained in the ensemble separately for target and nontarget scores. Column labels are the classifier labels from Table III.

	1	2	3	8	9	11	Avg.
ENet target	0.72	0.71	0.60	0.72	0.67	0.75	0.69
ENet non-target	0.61	0.61	0.55	0.54	0.44	0.48	0.55
Full ens. target	0.68	0.67	0.60	0.74	0.72	0.75	0.70
Full ens. non-target	0.55	0.56	0.55	0.57	0.53	0.54	0.56

In theory, elastic-net should, at least be equal to the best regularized fusion method, in all cases. But we notice that in the itv-itv condition, cross-validation selected $\alpha = 0$, instead of 1, as would have corresponded to the LASSO regularization. This will require further study on how to perform more accurate estimation of the α parameter.

Comparing the relative shrinkage factors $\hat{t}$, ridge $\geq$ LASSO $\geq$ elastic net. It is expected that ridge, as a non-sparse regularizer, shrinks less. Regarding the ensemble size, LASSO clearly retains the smallest number of base classifiers as expected. It is notable that, for the itv-itv case, LASSO zeroes half of the classifiers and achieves smallest error rates in all three metrics.

As a final analysis, Table VII, shows sparse fusion ensemble of six base classifiers that elastic net learned, namely $\{1, 2, 3, 8, 9, 11\}$. We show average pairwise Pearson's correlation, in a following way: for a fixed classifier i, we count $\frac{1}{|S|-1}\sum_{j \in S, j \neq i} \mathsf{Corr}(s_i, s_j)$, where S is the set of classifiers selected to an ensemble. Correlations were computed from non pre-warped scores. We count pairwise correlations between classifiers in the ensemble for target and non-target scores separately. In contrast, we also show average correlations in full ensemble to the selected base classifiers. We notice a slight reduction in average ensemble correlation from 0.70 to 0.69 for the targets and 0.56 to 0.55 for the non-targets. However, for classifiers 1 and 2, both target and non-target average correlations are increased. For the classifier 9, on the other hand, correlations are reduced from 0.72 to 0.67 and 0.53 to 0.44, for targets and non-targets respectively.

As we can see from the Table VIII, most correlated target score pairs are selected classifiers (1,2) and (8,11). Thus, maximum pairwise correlations non-selected classifiers are lower, for targets 0.80 to 0.78 and for non-targets 0.70 to 0.64. However, average maximum pairwise correlation in full-ensemble is still slightly bigger than the ENet ensemble. The fact that ENet considerably outperformed the full-ensemble, but still maximum pairwise correlations were not considerably reduced, is in line with the theoretical results proved in [43]. There, it was shown that pairwise correlations are not sufficient to predict ensemble accuracy, but that higher order correlations need to be considered.

TABLE VIII: As in Table VII, but using maximum pairwise correlation instead of the average.

	1	2	3	8	9	11	Avg.
Between selected classifiers							
ENet target	0.85	0.85	0.62	0.86	0.74	0.86	0.80
ENet non-target	0.83	0.83	0.73	0.63	0.55	0.63	0.70
From non-selected classifiers							
ENet target	0.77	0.75	0.69	0.82	0.82	0.82	0.78
ENet non-target	0.58	0.58	0.63	0.67	0.73	0.64	0.64
Full ens. target	0.85	0.85	0.69	0.86	0.83	0.86	0.82
Full ens. non-target	0.83	0.83	0.73	0.67	0.63	0.64	0.72

In summary, the analysis shows that while sparsity does indeed reduce pairwise correlation in the ensemble, correlation itself does not tell the full story of which classifers are redundant. One reason might be that pairwise correlation is unable to capture higher-order classifier dependencies. Similar observations have been made for instance in [43].

VI. Conclusion

We have presented a sparse regularized logistic regression score fusion for speaker verification. We optimized our system using audio data from NIST SRE 2008 corpus and evaluated using NIST SRE 2010 corpus (i.e. Evalset). We find that sparse regularization brings improvement over unregularized variant in all other sub-conditions and measures (EER, MinDCF, ActDCF) except in tel-tel condition.

In the condition itv-itv, LASSO regularization provided better performance than elastic-net. It shows that estimating the trade-off parameter α by cross-validation is not always successful. As a future work we plan to utilize Bayesian model selection techniques to automatically estimate both λ and α parameters from the fusion training set.

As a future work, it would be interesting to pursue methods that optimize ensemble diversity and ensemble classification error simultaneously as a way to obtain an ensemble with a good generalization property. Alternatively, run-time classifier ensemble selection for each speech utterance, similar to adaptable fusion using auxiliary quality measures would be an interesting direction.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *DSP*, vol. 10, no. 1, pp. 19–41, January 2000.

[3] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE T. Audio, Speech & Lang. Proc.*, vol. 16, no. 5, pp. 980–988, July 2008.

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

[6] C. P. Robert, *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, 2nd ed. Springer-Verlag, 2001.

[7] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, LLC, 2006.

[8] S. Pigeon, P. Druytsa, and P. Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *DSP*, vol. 10, no. 1–3, pp. 237–248, January 2000.

[9] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. Leeuwen, P. Matějka, P. Schwartz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE T-ASL*, vol. 15, no. 7, pp. 2072–2084, September 2007.

[10] L. Ferrer, K. Sönmez, and E. Shriberg, "An anticorrelation kernel for subsystem training in multiple classifier systems," *J. of Machine Learning Research*, vol. 10, pp. 2079–2114, 2009.

[11] N. Brümmer, "Fusion and toolkit [software package]," WWW page, June 2011, http://sites.google.com/site/nikobrummer/focal.

[12] "Bosaris toolkit [software package]," WWW page, June 2011, https://sites.google.com/site/bosaristoolkit/.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2000.

[14] V. Hautamäki, K. Lee, A. Larcher, T. Kinnunen, B. Ma, and H. Li, "Variational bayes logistic regression as regularized fusion for NIST SRE 2010," in *Speaker Odyssey 2012*, Singapore, June 2012.

[15] M. I. Jordan, "Why the logistic function? a tutorial discussion on probabilities and neural networks," Massachusetts Institute of Technology, Cambridge, MA, Tech. Rep., August 1995.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2008.

[17] G. Brown, A. Pocock, M. Zhao, and M. Luján, "Conditional likelihood maksimization: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[18] H. Li, B. Ma, K. A. Lee, H. Sun, D. Zhu, K. C. Sim, C. H. You, R. Tong, I. Kärkkäinen, C.-L. Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Tharmarajah, J. Epps, E. Ambikairajah, E.-S. Chng, T. Schultz, and Q. Jin, "The I4U system in NIST 2008 speaker recognition evaluation," in *ICASSP 2009*, Taipei, Taiwan, April 2009, pp. 4201–4204.

[19] F. Sedlák, T. Kinnunen, K. A. L. Ville Hautamäki, and H. Li, "Classifier subset selection and fusion for speaker verification," in *ICASSP 2011*, 2011.

[20] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, "The MITLL NIST LRE 2011 language recognition system," in *Speaker Odyssey 2012*, Singapore, June 2012, pp. 209–215.

[21] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, January 1992.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[23] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, August 2007.

[24] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[25] V. Hautamäki, K. Lee, T. Kinnunen, B. Ma, and H. Li, "Regularized logistic regression fusion for speaker verification," in *Interspeech 2011*, Florence, Italy, August, pp. 2745–2748.

[26] W. Campbell, D. Sturim, W. Shen, D. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition," in *Proc. ICASSP 2007*, vol. IV, 2007, pp. 217–220.

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

[28] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 3, pp. 2270–2285, 2005.

[29] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *DSP*, vol. 10, no. 1-3, pp. 42–54, January 2000.

[30] "Z-cal," 2006, http://ww.dsp.sun.ac.za/~nbrummer/focal/cllr/calibration/z_cal/index.htm. [Online]. Available: http://ww.dsp.sun.ac.za/~nbrummer/focal/cllr/calibration/z_cal/index.htm

[31] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, p. 308–313, 1965.

[32] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for L1 regularization: A comparative study and two new approaches," in *ECML 2007*, Warsaw, Poland, September 2007.

[33] A. Kabán, "On Bayesian classification with Laplacian priors," *Pattern Recognition Letters*, vol. 28, pp. 1271–1282, 2007.

[34] C.-L. Huang, H. Su, B. Ma, and H. Li, "Speaker characterization using long-term and temporal information," in *Proc. Interspeech 2010*, Makuhari, Japan, September 2010, pp. 370–373.

[35] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Sign. Proc. Lett.*, vol. 17, no. 6, pp. 599–602, 2010.

[36] D. Zhu, B. Ma, and H. Li, "Speaker verification with feature-space MAPLR parameters," *IEEE T-ASL*, vol. 19, no. 3, pp. 505–515, 2011.

[37] C. H. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE T-ASL*, vol. 18, no. 6, pp. 1300–1312, August 2010.

[38] D. van Leeuwen, "A note on performance metrics for speaker recognition using multiple conditions in an evaluation," Research note, June 2008, http://sites.google.com/site/sretools/cond-weight.pdf?attredirects=0.

[39] 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, May 2006.

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

[41] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development*. New Jersey: Prentice-Hall, 2001.

[42] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 957–968, 2005.

[43] G. Brown, "An information theoretic perspective on multiple classifier systems," in *Multiple Classifier Systems (MCS 2009)*, 2009, pp. 344–353.

Ville Hautamäki received the M.Sc. degree in Computer Science from the University of Joensuu, Finland in 2005. He received the Ph.D. degree in Computer Science from the same university in 2008. He has worked as a research fellow at the Institute for Infocomm Research, A*STAR, Singapore. Currently, he is post-doctoral researcher in University of Eastern Finland, funded by Academy of Finland. His current research interests are cluster analysis, speaker recognition and language recognition.

Tomi Kinnunen received the M.Sc. and Ph.D. degrees in computer science from the University of Joensuu, Finland, in 1999 and 2005, respectively. He worked as an associate scientist at the Speech and Dialogue Processing Lab of the Institute for Infocomm Research (I^{2}R), Singapore and as a senior assistant at the Department of Computer Science and Statistics, University of Joensuu, Finland. He is currently employed by the Academy of Finland as a post-doctoral researcher. His main research areas are speaker recognition and robust speech analysis. He has some interest to voice conversion and eye-movement analysis as well.

Filip Sedlák was born in 1985 in Brno, Czech Republic. He received his BSc degree from the Faculty of Information Technology (FIT) in Brno University of Technology (BUT). In 2008, he moved to University of Eastern Finland (UEF) as an exchange student. From February 2010 to July 2010, he was visiting Institute for Infocomm Research (I2R) and Nanyang Technological University (NTU), Singapore, as an intern. He is currently pursuing his master's degree in the School of Computing at UEF.

Kong-Aik Lee received his B. Eng. (first class honors) degree from University Technology Malaysia in 1999, and his Ph.D. degree from Nanyang Technological University, Singapore, in 2006. He is currently a senior research fellow with Human Language Technology department, Institute for Infocomm Research (I^{2}R), Singapore. His research focuses on statistical methods for speaker and spoken language recognition, adaptive echo and noise control, and subband adaptive filtering. He is the leading author of the book: *Subband Adaptive Filtering: Theory and Implementation*, Wiley, 2009.

Bin Ma (M'00-SM06) received the B.Sc. degree in Computer Science from Shandong University, China, in 1990, the M.Sc. degree in Pattern Recognition & Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, in 1993, and the Ph.D. degree in Computer Engineering from The University of Hong Kong, in 2000.

He was a Research Assistant from 1993 to 1996 at the National Laboratory of Pattern Recognition in IACAS. In 2000, he joined Lernout & Hauspie Asia Pacific as a Researcher focusing on the speech recognition of multiple Asian languages. From 2001 to 2004, he worked for InfoTalk Corp. Ltd as a Senior Researcher and a Senior Technical Manager engaging in mix-lingual telephony speech recognition and embedded speech recognition. Since 2004, he has been a Research Scientist and the Group Leader of Speech Processing group at the Institute for Infocomm Research, Singapore. He now serves as a Subject Editor of Speech Communication. His current research interests include robust speech recognition, speaker & language recognition, spoken document retrieval, natural language processing and machine learning.

Dr Haizhou Li is currently the Principal Scientist and Department Head of Human Language Technology at the Institute for Infocomm Research. He is also the Program Manager of Social Robotics at the Science and Engineering Research Council of A*Star in Singapore.

Dr Li has worked on speech and language technology in academia and industry since 1988. He taught in the University of Hong Kong (1988-1990), South China University of Technology (1990-1994), and Nanyang Technological University (2006-). He was a Visiting Professor at CRIN/INRIA in France (1994-1995), and at the University of New South Wales in Australia (2008). As a technologist, he was appointed as Research Manager in Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), and Vice President in InfoTalk Corp. Ltd (2001-2003).

Dr Li's research interests include automatic speech recognition, natural language processing and information retrieval. He has published over 150 technical papers in international journals and conferences. He holds five international patents. Dr Li now serves as an Associate Editor of IEEE Transactions on Audio, Speech and Language Processing, and Springer International Journal of Social Robotics. He is an elected Board Member of the International Speech Communication Association (ISCA, 2009-2013), a Vice President of the Chinese and Oriental Language Information Processing Society (COLIPS, 2009-2011), an Executive Board Member of the Asian Federation of Natural Language Processing (AFNLP, 2006-2010), and a Senior Member of IEEE since 2001. Dr Li was the local arrangement chair of SIGIR 2008 and ACL-IJCNLP 2009. He was the recipient of National Infocomm Award of Singapore in 2001. He was named one of the two Nokia Professors 2009 by Nokia Foundation in recognition of his contribution to speaker and language recognition technologies.