CLASSIFIER SUBSET SELECTION AND FUSION FOR SPEAKER VERIFICATION

Filip Sedlák[1], *Tomi Kinnunen*[1], *Ville Hautamäki*[2], *Kong-Aik Lee*[2], *Haizhou Li*[1,2]

[1]University of Eastern Finland, School of Computing, Joensuu, Finland
[2]Human Language Technology Deparment, Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{*fsedlak, tkinnu*}*@cs.joensuu.fi* {*vmhautamaki, kalee, hli*}*@i2r.a-star.edu.sg*

ABSTRACT

State-of-the-art speaker verification systems consists of a number of complementary subsystems whose outputs are fused, to arrive at more accurate and reliable verification decision. In speaker verification, fusion is typically implemented as a linear combination of the subsystem scores. Parameters of the linear model are commonly estimated using the logistic regression method, as implemented in the popular FoCal toolkit. In this paper, we study simultaneous use of classifier selection and fusion. We study four alternative fusion strategies, three score warping techniques, and provide interesting experimental bounds on optimal classifier subset selection. Detailed experiments are carried out on the NIST 2008 and 2010 SRE corpora.

Index Terms— Classifier selection, linear fusion

1. INTRODUCTION

Speaker verification is the task of accepting or rejecting an identity claim based on a person's voice sample [1]. Modern speaker verification systems utilize ensembles of *base classifiers* to arrive at an accurate and reliable verification decision by *classifier fusion*. The base classifiers might utilize, for instance, different speech parameterizations (e.g. spectral, prosodic, high-level), classifiers (e.g. Gaussian mixture models [2], support vector machines [3]) or channel compensation techniques (e.g. joint factor analysis [4], nuisance attribute projection [5]). For a given speech sample, each of the L base classifiers produces a *match score*, $s_i \in \mathbb{R}, i = 1, 2, \ldots, L$, that indicates the degree of belief for the target speaker hypothesis.

Having the set of base classifiers defined, the question remains how to combine the base classifier sub-decisions. In this paper, we restrict ourselves to *score-level* fusion $f : \mathbb{R}^L \rightarrow \mathbb{R}$ of the form $s = f(s_1, s_2, \ldots, s_L)$, where f is the fusion device that combines the L base classifier scores into a single match score, s. The binary accept/reject decision is then carried out by comparing the fused score to a pre-defined threshold, θ. In the system development phase, one uses a labeled training set to train the fusion parameters and the decision threshold. The trained fusion system f and threshold b are then used for making speaker verification decisions on unseen data. A natural goal, as in any pattern classification task, is to ensure good generalization on that unseen data.

There are three main affecting factors to the generalization performance, the choice of the fusion methodology (including threshold setting), score range normalization, and the choice of the base classifiers themselves. Regarding the fusion methodology, here we consider linear classifiers of the form $f_{\mathbf{w},\theta}(\mathbf{s}) = \mathbf{w}^t\mathbf{s} + \theta$ where $\mathbf{w} = (w_1, w_2, \ldots, w_L)^t$ is the vector of fusion weights, $\mathbf{s} = (s_1, s_2, \ldots, s_L)^t$ is the vector of base classifier scores and θ is the decision threshold; the speaker is accepted if and only if $f_{\mathbf{w},\theta}(\mathbf{s}) \geq 0$. The linear fusion scheme, when trained using the linear logistic regression objective in the *FoCal* toolkit[1], has been found robust and forms a good reference method. One of the successful elements of this method is warping of the fused scores into log-likelihood ratios (LLRs) prior to fusion. It is possible also to either precalibrate base classifier scores or postcalibrate fused scores [6].

The third factor affecting generalization performance, the choice of the base classifiers, is at least as critical as the fusion method. Whilst there does not currently exist a principled recipe for choosing the base classifiers, a general consensus is that the base classifiers errors should not be correlated [7]. One way to achieve this is to train base classifiers sequentially so that training of the current classifier takes into account errors produced by the previous classifers. This principle was applied to a fusion of two SVM-based classifiers with FoCal as a fusion method [8].

This paper represents our recent efforts in designing robust fusion strategies. During the latest NIST 2010 speaker recognition evaluation (SRE) benchmarking, the authors faced up a practical problem of fusing a dozen of spectral classifiers developed by independent sub-teams in our laboratories. One of the strategies considered, but not included in the SRE submission due to lack of time, was to use *classifier subset selection* together with linear fusion. The working hypothesis was that fusing a smaller number of (reliable) classifiers with fewer degrees of freedom would lead to more stable fusion. From a practical point of view, classifier selection would lead to computationally more feasible system as well. In addition to classifier selection, we studied extensively different score warping techniques and alternative ways to train linear fusion parameters, including an attempt for direct minimization of the NIST's decision cost function. The purpose of this paper is to summarize and conclude the lessons learned from these experiments.

2. CLASSIFIER FUSION AND SUBSET SELECTION

2.1. Problem Setup

We assume that, during the development phase, one has access to a development set $\mathcal{D} = \{(\mathbf{s}_i, y_i), i = 1, 2, \ldots, N_{\text{dev}}\}$ of base classifier score vectors $\mathbf{s}_i \in \mathbb{R}^L$, with $y_i \in \{+1, -1\}$ indicating whether the corresponding speech sample originates from a target speaker ($y_i = +1$) or from a non-target ($y_i = -1$). Given

The work of F. Sedlak was supported by the Nanyang Technological University (NTU), Singapore. The works of T. Kinnunen and V. Hautamäki were supported by Academy of Finland (projects 132129 and 131298 respectively). The work of H. Li was supported by Nokia foundation.

[1]`http://sites.google.com/site/nikobrummer/focal`

the linear fusion classifier $f_{\mathbf{w},\theta}(\mathbf{s}) = \mathbf{w}^t\mathbf{s} + \theta$ with parameters $(\mathbf{w}, \theta)$, and an empirical cost function $\mathcal{C}((\mathbf{w}, \theta), \mathcal{D})$, optimal fusion device is $(\mathbf{w}^{\text{dev}}, \theta^{\text{dev}}) = \arg\min_{(\mathbf{w},\theta)} \mathcal{C}((\mathbf{w}, \theta), \mathcal{D})$. Given another held-out dataset, $\mathcal{T} = \{(\mathbf{s}_i, y_i), i = 1, 2, \ldots, N_{\text{eval}}\}$, *actual cost* and *minimum cost* are computed as $\mathcal{C}((\mathbf{w}^{\text{dev}}, \theta^{\text{dev}}), \mathcal{T})$ and $\min_\theta \mathcal{C}((\mathbf{w}^{\text{dev}}, \theta), \mathcal{T})$, respectively. The difference of the actual and the minimum costs is known as *calibration error*.

In this study, we adopt the *decision cost function* (DCF) used in the NIST speaker recognition evaluations,

$$C_{\text{det}}(\theta) = C_{\text{miss}} P_{\text{miss}}(\theta) P_{\text{tar}} + C_{\text{fa}} P_{\text{fa}}(\theta)(1 - P_{\text{tar}}), \qquad (1)$$

where P_{tar} is the prior probability of a target (true) speaker, C_{miss} is the cost of a miss and C_{fa} is the cost of a false alarm. These application-dependent cost parameters can also be summarized as a single quantity known as *effective prior*: $P = \text{logit}^{-1}(\text{logit}(P_{\text{tar}}) + \log(C_{\text{miss}}/C_{\text{fa}}))$. It is possible to minimize DCF directly (e.g. [9]) or to optimize a proxy cost such as C_{wlr} [6]. As usual, we treat the fusion weights and the threshold independently; for each of the fusion training methods (Section 2.3) the threshold is set to $\theta^* = \arg\min_\theta C_{\text{det}}(\theta)$ after fusion training.

2.2. Score Pre-Warping

Since the base classifier scores s_i may have different interpretations (e.g. log-likelihood ratios or SVM inner products) and their scales may vary a lot, it is important to equalize their global range to avoid large-variance base classifier to dominate the fused score. We consider three such *score warping* methods, *mean and variance normalization* (MVN), *Z-calibration* (Z-cal) [6] and *S-calibration* (S-cal) [6]. Each of these methods includes a training phase to set its parameters.

The simplest method, MVN, normalizes the scores of each base classifier to have zero mean and unit variance. The parameters are the mean μ_i and standard deviation σ_i of the ith base classifier scores (determined from the development set). A given score s is warped according to $s' = (s - \mu_i)/\sigma_i$. Note that MVN is an unsupervised method that does not require class labels. The S-cal and Z-cal methods[2], in turn, are trained discriminatively by utilizing the target/nontarget key information of the development score vectors. They both aim at converting arbitrary scores to well-calibrated *log-likelihood ratios* (LLRs). The warping functions in S-cal and Z-cal are defined as,

$$\text{llr}_{\text{S-cal}}(s) = \log \frac{(\text{logit}^{-1}\alpha)(e^{xs+y} - 1) + 1}{(\text{logit}^{-1}\beta)(e^{xs+y} - 1) + 1},$$

$$\text{llr}_{\text{Z-cal}}(s) = (s - x_{\min})\frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}} + y_{\min}.$$

The parameters (slope x, offset y and saturation parameters α, β for S-cal; linear map parameters $x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$ for Z-cal) are trained through an iterative gradient-descent minimization of an LLR-based cost function [6].

2.3. Training Methods for Linear Fusion

In this paper, we consider four methods for linear fusion training. The first method, **equal weights**, performs simple averaging of the subsystems scores to produce the fused score, $f(\mathbf{s}) = \frac{1}{L}\sum_{i=1}^{L} s_i$. It has the advantage of notrequiring any training. The second method, **gradient C_{wlr} optimization**, uses an iterative discrete gradient descent method (we utilize MATLAB's `fminunc` function) to minimize the following effective-prior weighted log-likelihood ratio objective [6],

$$\text{C}_{\text{wlr}} = \frac{P}{N_t}\sum_{i=1}^{N_t} \log\left(1 + e^{-\mathbf{w}^t\mathbf{s}_i - \theta}\right) + \frac{1-P}{N_f}\sum_{j=1}^{N_f} \log\left(1 + e^{\mathbf{w}^t\mathbf{s}_j + \theta}\right), \qquad (2)$$

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. Here, P is the effective prior defined in subsection 2.1. In an initial stage of the study, we noticed that, for our base classifier scores, weight optimization in the FoCal toolkit did not always converge. Changing the optimization package solved the convergence problem.

The third method studied in this paper, **gradient MinDCF optimization**, is an attempt to minimize MinDCF (1) directly at a given operating point, rather than using the "soft" integration of all errors in (2). To mimic (2), we rewrite (1) as,

$$\text{C}_{\text{det}} = \frac{C_{\text{miss}} P_{\text{tar}}}{N_t}\sum_{i=1}^{N_t} g(-\mathbf{w}^t\mathbf{s}_i - \theta) + \frac{C_{\text{fa}}(1 - P_{\text{tar}})}{N_f}\sum_{j=1}^{N_f} g(\mathbf{w}^t\mathbf{s}_j + \theta),$$

where $g(x)$ is the unit step function, $g(x) = 1$ for $x \geq 0$ and $g(x) = 0$ elsewhere. To optimize (3), we use an iterative EM-like scheme as follows. We start with equal weights and find optimum θ as $\theta^* = \arg\min_\theta C_{\text{det}}(\theta)$. Having θ^* fixed, we optimize the weights in (3) using MATLAB's `fminunc` optimizer. The process is continued until convergence.

Finally, **greedy MinDCF optimization** is another method of direct MinDCF optimization. The weights are first initialized to be equal and then optimized one-by-one, in a greedy manner, by line search on $-1 \leq w_i \leq 1$. This method was used in our submission to the NIST 2010 SRE and is included here as a reference.

Table 1: Selection of the three datasets used in this study. We focus on the core-condition itv-tel subset with female trials.

Dataset	Usage	Data source	# Verif. trials
Training set	Train fusion params., do classif. selection	NIST 2008 itv-tel subset	263 t, 27315 f
Devset	Compare fusion and warping methods and classif. selection	NIST 2008 itv-tel subset	283 t, 27195 f
Evalset	Validate results	NIST 2010 itv-tel	801 t, 30254 f

2.4. Classifier Subset Selection

Up to this point, we have defined a standard fusion framework, assuming a full ensemble of L classifiers. Now, instead of optimizing the weights in the L-dimensional space, we are in a search of a globally optimum selection of $K \leq L$ classifiers, with K being unknown. For a fixed K, brute force requires search of $\binom{L}{K} = \frac{L!}{K!(L-K)!}$ possible classifier combinations, for which the fusion weights and decision threshold must be trained (note, however, that the score warping parameters need to be trained once only). Since K is also unknown, brute force requires evaluation of the entire powerset of $2^L - 1$ classifier fusions. This strategy, although not feasible in a realistic systems for large L, is chosen in the present study. Since we evaluate all the combinations, we can

[2] `http://www.dsp.sun.ac.za/~nbrummer/focal/cllr/calibration/`.

Table 2: Twelve base classifiers are constructed based on the four different cepstral features used in conjunction with four different speaker modeling techniques.

	Classifier	Feature	Devset (2008)		Evalset (2010)	
			EER (%)	MinDCF (x1000)	EER (%)	MinDCF (x1000)
1	GMM-UBM-JFA	PLP	**3.95**	0.7095	4.99	0.5547
2	GMM-UBM-JFA	PLP	4.24	0.6996	4.12	0.5267
3	GMM-UBM-JFA	PLP	4.24	0.6600	3.75	0.6840
4	GMM-UBM-JFA	LPCC	4.59	0.8735	5.74	0.7458
5	GMM-SVM-KL	PLP	5.65	0.6374	5.49	0.6522
6	GMM-SVM-KL	MFCC	4.99	0.5081	4.37	0.4955
7	GMM-SVM-KL	LPCC	6.45	0.5774	5.37	0.5954
8	GMM-SVM-KL	MLF	5.81	0.5590	4.74	0.5268
9	GMM-SVM-KL	LPCC	4.24	0.7158	6.52	0.6448
10	GMM-SVM-KL	SWLP	10.20	0.6897	5.87	0.5411
11	GMM-SVM-FT	PLP	8.13	0.6198	6.12	0.5872
12	GMM-SVM-BHAT	PLP	5.40	**0.4798**	**3.03**	**0.3371**

Table 3: Fusion of all 12 base classifiers on NIST SRE 2008 devset (itv-tel, females). First three rows show the best base classifiers.

Fusion method	Score warping	EER	MinDCF	ActDCF	ActDCF-MinDCF
Best ActDCF	–	4.99	0.5081	**0.8445**	0.3364
Best MinDCF	–	5.40	**0.4798**	0.9364	0.4566
Best EER	–	**3.95**	0.7095	0.9576	**0.2481**
Equal weights	–	2.47	0.3809	0.4402	0.0593
	MVN	2.47	0.3809	**0.4261**	0.0452
	S-cal	**2.20**	**0.3738**	0.4304	0.0566
	Z-cal	2.46	0.4310	0.4655	**0.0345**
Grad. C_{wlr}	–	**2.12**	**0.3477**	**0.3774**	0.0297
	MVN	**2.12**	**0.3477**	**0.3774**	0.0297
	S-cal	**2.12**	0.3498	0.3795	0.0297
	Z-cal	2.46	0.4016	0.4040	**0.0024**
Grad. MinDCF	–	2.47	0.4176	0.5257	0.1081
	MVN	2.36	0.4000	0.4925	**0.0925**
	S-cal	**1.77**	**0.3583**	**0.4664**	0.1081
	Z-cal	2.05	0.3876	0.4972	0.1096
Greedy MinDCF	–	2.83	0.4042	0.5822	0.1780
	MVN	**2.14**	**0.3498**	0.4876	0.1378
	S-cal	3.32	0.3618	**0.4699**	**0.1081**
	Z-cal	3.39	0.4860	0.5985	0.1125

be sure to have found the best possible classifier subset. Analogous to observing the difference of ActDCF and MinDCF costs, we can determine the best realizable classifier subset (found from training set) and compare the result to the optimum *oracle* selection on the evaluation set. In theory, classifier selection is simply a special case of linear fusion $f_{\mathbf{w},b}(\mathbf{s}) = \mathbf{w}^t\mathbf{s} + b$ where the weights of the excluded classifiers are set to zero. However, we have noted that in practice, the logistic regression optimization has difficulties to (completely) zero out the classifiers.

3. SPEECH CORPORA AND THE BASE CLASSIFIERS

We utilize the two most recent NIST SRE corpora, NIST 2008 and NIST 2010, for our experiments (Table 1). Due to space limits, we present all the results on the female[3] trials of the interview-telephone (int-tel) subcondition of the core tasks. The NIST 2008 trial list was split into two disjoint parts (with no speaker overlap), the first one used for training score warping, fusion weights and the decision threshold. The other part of the NIST 2008 data and the NIST 2010 data serve for validation purposes.

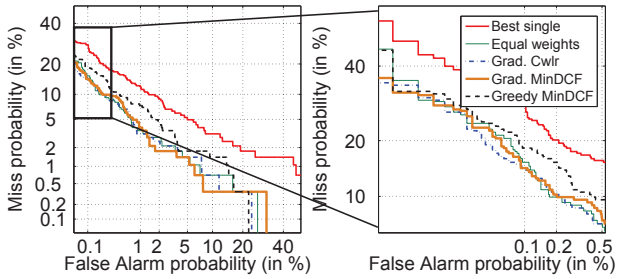

Fig. 1: Comparison of fusion methods using the full set of $L = 12$ classifiers (S-cal warping). The best individual classifier, in terms of ActDCF, is also displayed as a reference.

We use four speaker classification techniques in combination with four types of cepstral features in constructing the base classifiers. In particular, we used *perceptual linear prediction* (PLP), *linear predictive cepstral coefficient* (LPCC), *mel frequency cepstral coefficients* (MFCC), and the recently studied *stabilized weighted linear prediction* (SWLP) [10] features in parameterizing the speech utterances. Energy-based VAD was used to remove nonspeech frames. Additional RASTA filtering, cepstral mean/variance normalization (CMVN) and feature warping, were also applied.

Table 2 shows the twelve base classifiers based on four different cepstral features used in conjunction with four different classifiers. When two subsystems share the same classifier and features, it means that the systems are two independent implementations. For classifiers, we use the generative GMM-UBM-JFA [4] and the discriminative GMM-SVM approaches [11]. They are based on the universal background model (UBM) paradigm [2] and share similar form of subspace channel compensation, though the training methods differ. We used previous NIST SREs data, including SRE 2004, SRE 2005 and SRE 2006, to train the UBM and the session variability subspace. Switchboard data was also used 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 T-norm and Z-norm with SRE 2004 and SRE 2005 data as the background and cohort training data.

4. FUSION AND CLASSIFIER SELECTION RESULTS

We first compare the score warping and fusion training methods on the full set of $L = 12$ base classifiers in Table 3. As a reference, the first three rows display the best individual classifier per each considered cost function (EER, MinDCF, ActDCF). In addition, the last column shows the calibration error, $\mathrm{ActDCF} - \mathrm{MinDCF}$.

As expected, fusion improves accuracy over the best single classifier systematically. Regarding score warping, Z-cal yields systematically higher errors compared to MVN and S-cal (although Z-cal gives the smallest calibration errors in two cases). Comparing the fusion training methods (see also Fig. 1), the best EER (1.77 %) is

[3] This is a more difficult set than male trials.

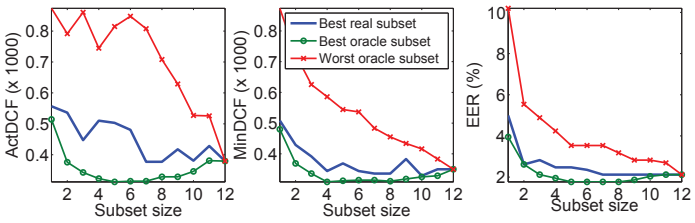

Fig. 2: Effect of classifier pool size to accuracy (NIST 2008 devset).

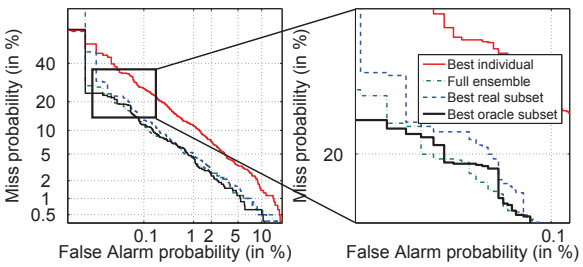

Fig. 3: Results on the NIST SRE 2010 evalset.

achieved, perhaps surprisingly, using the gradient MinDCF method. For MinDCF and ActDCF, however, gradient C_{wlr} leads to best result. Considering ActDCF and calibration error, the greedy MinDCF method yields the worst results, losing even to equal weights fusion.

We next study the issue of classifier selection on the NIST 2008 devset. Based on Table 3 and Fig. 1, we only consider gradient C_{wlr} method with S-cal. For a given subset size $K \leq L$ and performance metric (EER, MinDCF, ActDCF), we consider three summary values out from all the $\binom{12}{K}$ combinations in Fig. 2. The first one, *best real subset*, refers to optimum non-cheating classifier subset selection done on the training set and evaluated on the devset. The second value, *best oracle subset*, is computed by direct optimization on the devset, by knowing the key file. Note that the oracle selection considers only the subset selection – the fusion weights are still learnt from the training set. The third value, *worst oracle subset* refers to worst classifier subset selection on the devset and gives an idea how bad the result can be with unlucky classifier selection.

Is there any practical advantage of using a subset of classifiers instead of the full ensemble? Observing the middle (blue) lines in Fig. 2, especially for ActDCF, the answer seems *no*. The (green) oracle lines, however, reveal that there *does* exist a classifier subset which has potential to outperform full ensemble – only the prediction of that subset fails. Another interesting observation is that, for all three performance metrics, the empirical bounds on best and worst performance approach each other for increased subset size. This implies that, at least with these base classifiers, using a high number of classifiers leads to more stable fusion system, which is intuitively reasonable. Averaging "similar" spectral system scores, which are still independent, helps in reducing uncertainty of the fused score [12].

Table 4 summarizes the main results on NIST 2008 and NIST 2010 data, indicating best individual classifier, fusion of all classifiers, best non-cheating subset fusion and the best oracle subset. The subsets are searched from all the $2^{12} - 1 = 4095$ possible fusion combinations. The indices of the included classifiers are also shown. As seen, the optimum subsets include $K = 5$ classifiers (it is a coincidence that the real and oracle subset sizes are equal). It is interesting to note that only *one* classifier (no 6) is common to the real and oracle sets. The DET curve for the NIST 2010 data in Fig. 3 agrees with the results in Fig. 2. That is, although the real subset selection fails to improve over the full ensemble, the oracle suggests that there is room for improvement, especially at low false alarm rates.

Table 4: Comparing different classifier subset selections. The best individual classifier is selected according to ActDCF.

	Fusion	Included classifiers	EER (%)	MinDCF (x1000)	ActDCF (x1000)
Devset (2008)	Best individual	12	5.40	0.4798	0.5144
	Full ensemble	all	2.12	0.3498	0.3795
	Best real subset	{1,2,3,4,6}	2.51	0.3689	0.5031
	Best oracle subset	{5,6,7,9,11}	3.18	0.3124	0.3124
Evalset (2010)	Best individual	2	4.12	0.5267	0.5561
	Full ensemble	all	2.58	0.3089	0.3661
	Best real subset	{1,2,3,4,6}	2.45	0.3644	0.7019
	Best oracle subset	{1,4,6,8,10}	2.22	0.2715	0.2740

5. CONCLUSION

The results confirm the view that combination of S-cal and linear fusion training with logistic regression training performs well. Fusion, in general, leads to remarkable improvements over the best individual classifier. We found that increasing classifier pool size yields more reliable fusion as compared to using a subset. Although classifier subset selection did not improve accuracy over the full ensemble, the oracle result indicated potential of the method. It would be therefore interesting to study if the well-fusing classifier subsets could be better predicted either at the development phase, or during runtime.

6. 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.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[3] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. 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.M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP 2005*, Philadelphia, Mar. 2005, pp. 629–632.

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

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

[8] 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.

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

[10] 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.

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

[12] N. Poh and S. Bengio, "Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks?," in *Proc. ICASSP 2004*, Montreal, Canada, May 2004, vol. 5, pp. 893–896.