SPEAKER RECOGNITION USING MATCHED FILTERS

Hagai Aronowitz

IBM Research - Haifa
hagaia@il.ibm.com

ABSTRACT

Nowadays state-of-the-art speaker recognition systems obtain quite accurate results for both text-independent and text-dependent tasks as long as they are trained on a fair amount of development data from the target domain, and as long as the target data is clean. In this work we investigate the use of matched filters for speaker recognition in the framework of a small in-domain development data. We show how a matched filter can be optimized to maximize SNR (signal to noise ratio) when the noise component includes both intra-speaker variability and center/mean hyper-parameter variability. The proposed method generalizes our previous method named score stabilization and obtains significant speaker recognition error reductions.

Index Terms— speaker verification, domain adaptation, score normalization, i-vector centering, matched filter

1. INTRODUCTION

The introduction of i-vectors [2] and Probabilistic Linear Discriminant Analysis (PLDA) [3] resulted in very low error rates in the recent NIST text-independent (TI) speaker recognition evaluations (SREs) [4]. However, the success of i-vector based PLDA is dependent on the availability of a large development set with thousands of multi session speakers, to estimate the PLDA hyper-parameters. Moreover, the development data must be matched to the target data.

When the target data is highly mismatched to the available development data, for instance due to channel mismatch or in text-dependent speaker recognition, a common strategy is to collect some data from the target domain. The collected in-domain data is then used to either train the speaker recognition system from scratch [4-7] or to adapt an already existing system [5-10].

In this paper we investigate how to train a speaker recognition system with limited in-domain data (no use of out-of-domain data whatsoever). We assume that each session is parameterized by a high-level feature vector (such as an i-vector or a supervector), and propose to estimate a matched filter for each enrolled speaker. The speaker-dependent matched filter is then used to produce scores. We then apply standard score normalization (ZT-norm [11]).

The use of a matched filter for scoring high level features in the framework of speaker recognition has been proposed in [12] for suppressing an interfering speaker in the framework of speaker recognition in summed (2-wire) conversations. The matched filter was applied in the Gaussian Mixture Model (GMM) supervector space.

Variants of matched filters were successfully used by Khosravani [13] for training a speaker recognizer with unlabeled development data (in the framework of the NIST 2013-4 *i-vector challenge*).

The use of matched filters in this work is different in the sense that the focus is training the system using a small labeled development dataset. Furthermore, the noise/interference we aim at suppressing is a combination of intra-speaker inter-session variability and center/mean hyper-parameter variability.

The center/mean hyper-parameter is the expectation of the distribution of high-level features (i-vectors, GMM or DNN supervectors) over the target session space. In the i-vector LDA/PLDA framework the mean is used for i-vector centering prior to length normalization. The importance of having an accurate estimation of the i-vector center has been highlighted in [8]. For the GMM-NAP (Nuisance Attribute Projection) framework score-normalization implicitly uses an estimate of the center/mean [11], and the importance of having an accurate estimation of it has been shown in [1]. In practice, when devset size is limited, the point estimate of the center/mean hyper-parameter is noisy and this is addressed by our proposed method.

The remainder of this paper is organized as follows: Section 2 describes the proposed method. Section 3 describes the baseline system, data, experiments and results. Finally, Section 4 concludes.

2. MATCHED FILTERS FOR SPEAKER RECOGNITION

2.1. Matched filter

We assume that an observed signal x is a sum of a desirable signal s and an additive noise v:

$$x = s + v \tag{1}$$

We seek a filter h, such that h maximizes the output signal-to-noise ratio, where the output is the inner product of the filter and the observed signal x. The solution is given by [14]

$$h = \frac{1}{\alpha}\left(\mathrm{cov}\{v\}\right)^{-1} s \tag{2}$$

where α is a scaling constant dependent on s and $\mathrm{cov}\{v\}$ (α cancels out after score normalization).

2.2. Model

Given a high-level feature vector x extracted from a session, consider the following generative model:

$$x = c_x\left(s + n_x - \mu\right) + \mu \tag{3}$$

where s is the mean high-level feature vector representing the speaker, n_x is a session dependent intra-speaker nuisance vector, μ is the center/mean of the speaker population distribution (not necessarily known) of high-level features, and c_x is a session dependent scaling factor. The scaling factor is required to model a scaling phenomenon which is partly the basis of popular methods such as score normalization, cosine distance scoring, and i-vector length normalization.

Given a pair high-level features x and y corresponding to an enrollment session and a verification session respectively, consider scoring function $\boldsymbol{f}$:

$$f(x,y)=(x-\mu)^t\Omega(y-\mu) \tag{4}$$

where Ω is a matrix we strive to optimize. Scores are then normalized using ZT-norm:

$$\begin{aligned} f_{Znorm}(x,y) &= \frac{f(x,y)}{\sqrt{\mathrm{var}_{x'}\{f(x',y)\}}} \\ f_{ZTnorm}(x,y) &= \frac{f_{Znorm}(x,y)}{\sqrt{\mathrm{var}_{y'}\{f_{Znorm}(x,y')\}}} \end{aligned} \tag{5}$$

Note that Eq. (5) is free of score bias terms because they biases equal to zero for f, as can be seen in Eq. (6). The variance terms in Eq. (5) are estimated from the (small) development dataset.

$$\mathrm{E}_{x'}\{f(x',y)\}=\left(\mathrm{E}_{x'}\{(x'-\mu)\}\right)^t\Omega(y-\mu)=0 \tag{6}$$

In practice, μ is unknown and is replaced by an estimate. We use our estimate to center our vector space (remove the estimated center from each vector in our development and evaluation data). We denote the estimate bias (error) by vector δ. $\tilde{x}$ denotes a centered vector associated to speaker s. $\tilde{x}$ can be reformulated as:

$$\tilde{x}=c_x\left(s+n_x+\frac{\delta}{c_x}\right) \tag{7}$$

Our scoring framework includes ZT-norm which is invariant to scaling of the input vectors. Therefore we can replace for simplicity Eq. (7) with Eq. (8):

$$\tilde{x}=s+n_x+\frac{\delta}{c_x} \quad . \tag{8}$$

Centered vector $\tilde{y}$ originating from the same speaker as $\tilde{x}$ can now be formulated as:

$$\tilde{y}=s+n_y+\frac{\delta}{c_y}=\tilde{x}+n_y-n_x+\delta\left(\frac{1}{c_y}-\frac{1}{c_x}\right) \tag{9}$$

2.3. Suppressing target variability

The matched filter corresponding to the model in Eq. (9) is

$$h=\tfrac{1}{\alpha}\left(\mathrm{W}+\mathrm{var}\{\tfrac{1}{c}\}\Delta\right)^{-1}\tilde{x} \tag{10}$$

where W stands for the intra speaker covariance matrix

$$\mathrm{W}=\mathrm{cov}\{n\} \tag{11}$$

and Δ stands for the center/mean uncertainty covariance matrix

$$\Delta=\mathrm{cov}\{\delta\} \tag{12}$$

Note that a reasonable estimate for Δ is based on the sample total covariance matrix T (estimated from the development set)

$$\Delta\cong\frac{\mathrm{T}}{m} \tag{13}$$

where m is the number of speakers in the development dataset (see [1] for more details). Note also that the effect of center uncertainty is magnified by the scaling variability.

2.3. Scoring with the matched filter

Once h is estimated according to Eq. (10). we set Ω (in Eq. (4)) to be equal to $\left(\mathrm{cov}\{v\}\right)^{-1}$ (in Eq. (2)). That is, we obtain the scoring function

$$f(x,y)=\tilde{x}^t\left(\mathrm{W}+\frac{\mathrm{var}\{\frac{1}{c}\}}{m}\mathrm{T}\right)^{-1}\tilde{y} \tag{14}$$

for suppressing target variability. The scoring function in Eq. (14) is evaluated in Section 3.

2.5. Implementation issues

2.5.1. Smoothing

W and T are estimated from a small development set. Therefore, they are not invertible and noisy. We smooth both W and T by using the shrinkage method [15].

2.5.2. NAP vs. WCCN (Within Class Covariance Normalization)

Assuming an infinite development dataset (m→∞), Eq. (14) turns into

$$f(\tilde{x},\tilde{y})=\tilde{x}^t\mathrm{W}^{-1}\tilde{y} \tag{15}$$

which turn out to be the WCCN method [16]. However from our past experience (revalidated on the setup described in Section 3), NAP (which is hard subspace removal for intra-speaker variability compensation) slightly but consistently outperforms WCCN for the GMM-supervector framework. We therefore apply NAP as a preprocessing step on vectors x and y, and set W (which is now the residual intra-speaker variability) to be a scalar matrix. The value of the scalar is estimated from the development data.

2.5.3. Estimating $\mathrm{var}\{\frac{1}{c}\}$

We estimate $\mathrm{var}\{\frac{1}{c}\}$ from the development data. We assume for this purpose that $\delta=0$ and $n=0$. We compute the mean high-level vector for each speaker in the development set and estimate the scaling factor c_x for each session x with respect to the speaker mean. Finally we estimate $\mathrm{var}\{\frac{1}{c}\}$ as the empirical variance of $\{\frac{1}{c_x}\}$ over the development dataset.

3. EXPERIMENTS

3.1. Baseline System

Our baseline system is based on the GMM-NAP framework as GMM-NAP outperforms i-vector based approaches when development data is small [6, 7]. Nevertheless, our proposed method can be used for the i-vector framework as well to handle center uncertainty.

In the GMM-NAP framework a GMM is adapted for each session (enrollment, testing and development) from a UBM using MAP-adaptation. A projection is estimated from the development set and is used to compensate intra-speaker intersession variability (such as channel variability).

3.1.1. Front-end

The front-end is based on Mel-frequency cepstral coefficients (MFCC). An energy based voice activity detector is used to locate and remove non-speech frames. The final feature set consists of 12 cepstral coefficients augmented by 12 delta and 12 double delta coefficients extracted every 10ms using a 25ms window. Feature warping is applied with a 300 frame window before computing the delta and double delta features.

3.1.2. GMM supervector extraction

A 512-Gaussian Universal Background Model (UBM) with diagonal covariance matrices is trained on the development set and is used for extracting the supervectors. The means of the GMMs are stacked into a supervector after normalization with the corresponding standard deviations of the UBM and multiplication by the square root of the corresponding weight from the UBM:

$$x = \Sigma^{-1/2}\left(\lambda_{UBM}^{1/2} \otimes I_F\right)\mu \tag{16}$$

where μ stands for the concatenated GMM means, λ_{UBM} stands for the vectorized UBM weights, $\sum$ is a block diagonal matrix with covariance matrices from the UBM on its diagonal, F is the feature vector dimension, $\otimes$ is the Kronecker product, and I_F is the identity matrix of rank F. We center all supervectors using the mean of the development set.

3.1.3. NAP estimation

A low rank projection P is estimated as follows. First, we remove from each supervector in the development its corresponding speaker supervector mean. The resulting supervectors are named nuisance supervectors. We compute the covariance matrix of the nuisance supervectors and apply PCA to find a basis to the nuisance space. Projection P is created by stacking the top k eigenvectors as columns in matrix V:

$$P = I - VV^t. \tag{17}$$

3.1.4. NAP compensation

The enrollment supervectors are compensated by applying projection P.

$$x_{compensated} = Px. \tag{18}$$

3.1.5. Scoring and score normalization

Scoring is performed using a dot-product between the compensated enrollment and test supervectors. We apply ZT-score normalization [11] using the sessions from development data.

3.2. Contrasting System: Score stabilization

In [1] we aimed at improving our GMM-NAP system in the small development dataset scenario. We proposed to stabilize score normalization parameters by removing from the GMM-supervector space a subspace spanned by the top eigenvectors of the total variability covariance matrix (hence, *score stabilization*).

In fact, our proposed scoring function (Eq. (14)) replaces the hard subspace removal in [1] by a soft approach which effectively deemphasizes the top eigenvectors of the total variability covariance matrix.

3.3. Text dependent dataset

The WF dataset consists of 750 speakers which are partitioned into a development set (200 speakers) and an evaluation dataset (550 speakers). Each speaker has 2 sessions using a landline phone and 2 sessions using a cellular phone. The data collection was accomplished over a period of 4 weeks.

In this work we limit ourselves to the common passphrase condition for which the same passphrase is used for both development, enrollment and verification. We report results for the 10-digit pass phrase 0-1-2-3-4-5-6-7-8-9 which we name ZN.

In the WF dataset each session contains 3 repetitions of ZN. For each enrollment session we use all 3 repetitions for enrollment, and for each verification session we use only a single repetition. A comprehensive description of the WF dataset can be found in [4].

We define the following subsets of the WF dataset (Table 1). In table 1 **L** stands for a landline sessions and **C** for a cellular session. For instance, **LLCC** stands for 4 sessions (2 landline + 2 cellular), and **LC** stands for 2 sessions (1 landline + 1 cellular). Subsets are gender balanced.

Table 1. Reduced development dataset.

Name	Number of speakers	Sessions per speaker
Full	200	LLCC
50	50	LLCC
50LC	50	LC
30	30	LLCC
30LC	30	LC
30LL	30	LL
30CC	30	CC
20	20	LLCC
20LC	20	LC

3.4. Text independent dataset

We use the NIST 2010 SRE [10] for evaluation. We use the NIST 2010 SRE male core trial list with telephone conditions (5, 6 and 8) for evaluation. The dataset consists of 355, 178 and 119 target trials and 13746, 12825 and 10997 impostor trials respectively.

The development dataset consists of male sessions from NIST 2004 and 2006 SREs (telephone data only). In total we use 4374 sessions from 521 speakers.

tWe define the following subsets of the TI development set. The number of speakers is varied between 20 and 500. Each subset consists of 2 sessions per speaker.

3.5. Text dependent results

Table 2 reports results using different subsets for development. The baseline system (with NAP subspace dimension of 10 which was found optimal in [1]) is contrasted to both score stabilization-based system and to the proposed method. In order to reduce the variance of our measured EERs, we repeat each experiment 10 times with randomly selected subsets. the score stabilization-based system is configured to the best configuration found in [1] (removal of top 25 eigenvectors of the total variability covariance matrix).

Table 2. Results for TD using different subsets for development. The proposed method is contrasted to both the baseline and score stabilization (SS) systems. Results are averaged over 10 randomly selected subsets. Best result for each subset is in bold.

System	20LC	20	30CC	30LL	30LC	30	50LC	50	Full
Baseline	2.8	2.5	3.2	3.3	2.4	2.1	1.8	1.6	1.0
SS	2.3	2.0	2.4	2.4	2.1	1.8	1.7	1.5	1.1
Matched Filter	**2.2**	**1.9**	**2.3**	**2.3**	**1.9**	**1.7**	**1.6**	**1.4**	**0.9**
Error reduc. rel. to baseline in %	21	24	28	30	21	19	11	13	10
Error reduc. rel. to SS in %	4	5	4	4	10	6	6	7	18

3.6. Text independent results

Table 3 reports results using different subsets of the development dataset. The baseline system (with NAP subspace dimension of 50 which was found optimal in [1]) is contrasted to both the contrasting system and to the proposed method. In order to reduce the variance of our measured EERs, we repeat each experiment 10 times with randomly selected subsets. the contrasting system is configured to the best configuration found in [1] (removal of top 10 eigenvectors of the total variability covariance matrix).

Table 3. Results for the TI task as a function of number of speakers in subset. Subsets contain **two** sessions per speaker. Results are averaged over 10 randomly selected subsets. Best result for each subset is in bold.

Method	Cond.	20	30	40	50	100	200	300	400	500
Baseline	5	13.7	13.2	11.8	10.7	9.6	7.3	6.2	5.4	5.4
SS		13.7	13.0	**11.0**	10.1	**8.7**	**6.5**	**5.6**	**4.5**	**5.1**
Matched Filter		**13.5**	**12.7**	**11.0**	**9.8**	8.9	6.8	5.9	5.1	5.3
Baseline	6	16.3	14.7	14.7	14.5	14.0	9.6	8.4	7.9	**7.3**
SS		15.1	15.0	14.6	**13.5**	**11.8**	**8.3**	**8.2**	**7.3**	**7.3**
Matched Filter		**14.9**	**14.6**	**14.2**	**13.5**	**11.8**	9.1	7.9	7.5	**7.3**
Baseline	8	6.7	5.9	5.0	5.0	4.2	2.5	**1.7**	**1.5**	**1.7**
SS		6.7	6.7	**4.2**	**4.1**	**1.7**	**1.7**	**1.7**	1.7	**1.7**
Matched Filter		**5.9**	**5.2**	5.0	4.2	2.6	**1.7**	**1.7**	1.7	**1.7**

4. CONCLUSIONS

In this work we generalize our recently proposed method of coping with uncertainty in center and total variability covariance matrix estimate. Contrary to our method in [1] which removes a subspace spanned by the top eigenvectors of the total variability covariance matrix, we use a softer approach of using the matched filter framework to optimally suppress the combination of the intra-speaker variability and center/mean uncertainty.

The proposed method was evaluated under the GMM-NAP framework as it has been found in the past to outperform the i-vector framework when development data is limited [4-7].

For the text dependent experiments, the proposed method improves significantly over the baseline (by 20% relative in average) and over the score stabilization method (by 7% relative in average). For the text independent experiment, the proposed method outperforms the baseline in almost all experiments, and outperforms the score stabilization method for smaller amounts of speakers in the development set.

The superior results for score stabilization for larger amounts of speakers may hint that our proposed method should suppress more aggressively the total variability covariance matrix, probably to suppress other sources of variability, for instance, inter-imposter variability as been done in [13]. Note also that contrary to score stabilization, our proposed method does not directly suppress the noisy estimates of the variance parameters in score normalization.

5. ACKNOWLEDGEMENTS

This work was part of the DEM@CARE EU project, partly funded by the European Commission in the scope of the 7th ICT framework.

12. REFERENCES

[1] H. Aronowitz, "Score Stabilization for Speaker Recognition Trained on a Small Development Set", in Proc. *Interspeech*, 2015.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," IEEE *Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788 - 798, 2010.

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in Proc. *Interspeech*. 2011.

[4] H. Aronowitz, R. Hoory, J. Pelecanos, D. Nahamoo, "New Developments in Voice Biometrics for User Authentication", in Proc. *Interspeech*, 2011.

[5] H. Aronowitz, "Text Dependent Speaker Verification Using a Small Development Set", in Proc. *Speaker Odyssey*, 2012.

[6] H. Aronowitz, O. Barkan, "On Leveraging Conversational Data for Building a Text Dependent Speaker Verification System", in Proc. *Interspeech*, 2013.

[7] H. Aronowitz, A. Rendel, "Domain Adaptation for Text Dependent Speaker Verification", in Proc. *Interspeech*, 2014.

[8] H. Aronowitz, "Inter dataset Variability compensation for speaker recognition", in Proc. *ICASSP*, 2014.

[9] H. Aronowitz, "Compensating Inter-Dataset Variability in PLDA Hyper-Parameters for Robust Speaker Recognition", in Proc. *Speaker Odyssey*, 2014.

[10] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised Domain Adaptation for i-vector Speaker Recognition," in Proc. *Speaker Odyssey*, 2014.

[11] H. Aronowitz, V. Aronowitz, "Efficient score normalization for speaker recognition", in Proc. *ICASSP*, 2010.

[12] H. Aronowitz and Y.A. Solewicz, "Speaker Recognition in Two Wire Test Sessions," in Proc. *Interspeech*, 2008.

[13] A. Khosravani, M.M. Homayounpour, "Linearly Constrained Minimum Variance for Robust I-vector Based Speaker Recognition ", in Proc. *Speaker Odyssey*, 2014.

[14] G. L. Turin, "An introduction to matched filters." IRE Transactions on Information Theory 6 (3) (1960): 311- 329.

[15] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices", Journal of Multivariate Analysis, 2004.

[16] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class Covariance Normalization for SVM-based Speaker Recognition," Proc. of *Interspeech*, 2006.