Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Reducing Noise Bias in the i-Vector Space for Speaker Recognition

Yosef A. Solewicz[1], *Hagai Aronowitz*[2], *Timo Becker*[3]

[1] Technology Section, Israel National Police, Jerusalem, Israel
[2] IBM Research – Haifa, Israel
[3] Federal Criminal Police Office, Germany

solewicz@police.gov.il, hagaia@il.ibm.com, timo.becker@bka.bund.de

Abstract

In this paper we develop a simple mathematical model for reducing speaker recognition noise bias in the i-vector space. The method was successfully tested on two different databases covering distinct microphones and background noise scenarios. Substantial reduction in score variability was attained across distinct evaluation conditions which is particularly important in forensic applications. Although originally designed for addressing additive noise, we show that under certain circumstances the proposed method incidentally alleviates convolutive nuisance as well.

Index Terms: speaker recognition, forensics, noise robustness, score compensation

1. Introduction

It is widely known that noise is one of the factors which significantly contribute to performance degradation in speaker recognition. Generally, noise can be categorized either as additive or convolutive. While the recording environment is usually responsible for the additive noise, convolutive noise is rooted on peculiarities of distinct microphones and transmitting channels.

Nuisance issues are presently being accentuated with the advent of numerous models of smartphones and mini-tape recorders widely disseminated in the market. The prompt availability of these devices leads to the production of recordings accompanied by a countless combination of microphone and background nuisance. This scenario poses a serious challenge to applications demanding calibrated recognition scores such as in forensics. In this regard, collecting specific reference populations matching this multiplicity of model\test nuisance patterns is impractical. Methods for compensating score bias are required to alleviate the requests of precise reference data reflecting each specific recording scenario.

State-of-the-art speaker recognition is based on mapping the speech signals into a high-level feature vector space, followed by the application of linear compensation and then computing vector similarity. Examples of such high-level vector representations are GMM supervectors and i-vectors [1]. Research efforts have been focusing on reducing the effect of handset/channel mismatch. Compensation techniques addressing the feature, model and score domains have been proposed [2]. In particular, techniques such as Nuisance Attribute Projection (NAP) [3], Joint Factor Analysis [4] and Probabilistic Linear Discriminant Analysis (PLDA) [5] were developed to address this problem.

In this paper, we propose a model for reducing the recognition score bias in the i-vector domain that can be applied as a supplementary noise compensation layer. The method relies on non-speech portions of the signal to estimate noise i-vectors in order to predict the noise impact on the i-vector space. The recognition score is then compensated using the estimated bias. Note that our method differs from common data-driven approaches (such as [6]) as we do not explicitly denoise the i-vectors.

The paper is organized as follows. In Session 2, we develop the theoretical foundations of the model. In Session 3, we report validation experiments. Session 4 suggests a novel quality measure for speech signals directly inspired by the proposed approach. Conclusions and future research are presented in Session 5.

2. Model

When speech is captured by some recording device, the signal is modified by the transmission channel between talker and microphone. The channel is a general entity fusing several factors such as the effect of the acoustic environment, microphone type and location, and recording device settings.

In addition, there may be further degradations by the background noise. The whole process may be expressed as

$$Y(t) = H(t) * X(t) + N(t) \quad (1)$$

where $Y(t)$ and $X(t)$ are the observed and clean speech signals respectively, $H(t)$ is a filter representing the convolutive slowly-varying channel effects, $N(t)$ denotes the additive background noise, and the ***** operator represents convolution. The observed signal is somewhat distorted and clearly impacts intelligibility and speech/speaker recognition applications.

In particular, it can be shown [7] that the cepstrum of the corrupted speech can be modeled by the sum of the clean speech, channel filter and (a function of the) additive noise cepstra. In this paper, however, we circumvent explicit estimation of the noise components, and ultimately tackle the score bias due to underlying noise effects at the i-vector space.

Our proposed model described below directly addresses additive noise components (N) although convolutive nuisance (H) is incidentally approached as well. The fact that channel characteristics are still being modeled by background noise can be understood as follows [8]. The non-speech spectrum can be modeled as the sum of the spontaneous activity of the channel and the transmitted background noise. Without further knowledge about these two components, non-speech portions of the signal cannot be used to obtain a reliable estimation of the channel transfer function itself, although they still capture channel information to some extent.

Assuming proper channel compensation, we redefine (1), eliminating $H(t)$. The observed signal is expressed as follows.

doi: 10.21437/Odyssey.2016-54

$$Y(t) = X(t) + N(t)\ . \quad (2)$$

We define $\boldsymbol{\alpha}$ (a kind of inverse signal-to-noise ratio (SNR)), to be the ratio of the standard deviations of the noise amplitude and the observed signal amplitude (not the clean signal amplitude):

$$\alpha = \frac{\sigma(N)}{\sigma(Y)}\ . \quad (3)$$

In the i-vector framework, *Y*, *X* and *N* are mapped into *y*, *x* and *n* respectively. We approximate the observed i-vector *y* as a linear function of the clean i-vector *x*; the noise i-vector *n*, estimated on non-speech frames of *y*; and α. Note that for the clean condition (α=0) we obtain *y*=*x*; for an extremely noisy condition (α=1) we obtain *y*=*n*; and for a SNR of 0 dB (α=0.5) we obtain *y*=0.5*x*+0.5*n*. Our model therefore assumes that the observed i-vector *y* is the following weighted average between the desired i-vector *x* and the noise i-vector *n*:

$$y = (1-\alpha)x + \alpha \cdot n. \quad (4)$$

We use the cosine distance to measure the similarity between i-vectors. The cosine distance between observed i-vectors y_1 and y_2 is given by

$$\langle y_1, y_2 \rangle = \frac{y_1 \cdot y_2}{|y_1||y_2|}. \quad (5)$$

where • represents the dot product between y_1 and y_2. Using (4) and after some algebra, we can express the similarity between desired vectors x_1 and x_2 in terms of the observed similarity $\langle y_1, y_2 \rangle$ and the respective noise i-vectors n_1 and n_2, with noise levels α_1 and α_2:

$$\langle x_1, x_2 \rangle = c_1 \langle y_1, y_2 \rangle - c_2 \quad (6)$$

where

$$c_1 = \frac{1}{(1-\alpha_1)(1-\alpha_2)} \quad (7)$$

is a multiplicative bias term that compensates for the drop in variance of the signals due to the added noise and

$$c_2 = \frac{\alpha_1 n_1 \cdot y_2 + \alpha_2 n_2 \cdot y_1 - \alpha_1 \alpha_2 n_1 \cdot n_2}{(1-\alpha_1)(1-\alpha_2)|y_1||y_2|} \quad (8)$$

is the additive bias component. $\boldsymbol{c_2}$ itself can be further parsed into a cross-term relating the observed signals and their background noise signals and a diagonal term tying the two noise signals. (Note that our method is invalid when $\boldsymbol{\alpha_1}$ or $\boldsymbol{\alpha_2}$ is close to unity).

We assume that $|\boldsymbol{y}| \approx |\boldsymbol{x}|$ since we target setups with relatively high SNRs where α is small and *x* and *n* are fairly orthogonal, implying that *y* should be a slightly rotated version of *x*.

3. Experiments

The experiments performed in this work are based on two databases, each one focusing on either additive or convolutive noise. The main database covering essentially the additive noise scenario is the DAPS (Device and Produced Speech) Dataset [9], which is a collection of studio speech recordings on tablet and smartphone devices in real world environments. The second database is a collection of several distinct microphone recordings extracted from the NIST 2005 Speaker Recognition Eval (SRE) [10], thus mainly focusing on channel nuisance (though additive noise is also an issue).

3.1. Protocols

3.1.1. DAPS

The DAPS dataset consists of 10 male speakers reading 5 excerpts each from public domain books. The recordings were done in a professional recording studio. Multiple versions of the data were created from these initial recordings. In the first version, a professional sound engineer applied audio effects to create production quality speech. In the other versions, the initial recordings were played through a high quality loudspeaker in real world environments and recorded onto one of three consumer devices, yielding a total of 12 device-noise conditions. These conditions are listed in Table 1 below. We create 2600 target trials and 23400 impostor trials, where enrollment is done on the original studio recordings, and testing is equally distributed among the distinct produced conditions.

3.1.2. NIST

The NIST benchmark is based on 1200+ recordings from 46 speakers in eight auxiliary microphones extracted from the NIST 2005 SRE. We benchmarked each of the microphones, trialing its recordings (models) against the other microphone recordings (tests), leading to distinct conditions containing about 6000 trials equally distributed between targets and impostors.

3.2. System

The speaker recognition system used in these experiments is an i-vector based system. The feature set consists of 13 Mel-cepstral coefficients appended by their differentiates. Sufficient statistics for i-vector extraction are derived from a 512 dimensional GMM. Following a 400-length i-vector extraction, linear discriminant analysis (LDA) is applied [1], compressing the i-vector dimension to 200. (No further channel compensation operations such as WCNN are applied.) Voice activity detection was performed using a phoneme recognizer [11] and scoring is implemented through cosine distance between model and test vectors, as discussed. The system was trained with telephone data from NIST 2004, 2006 and 2008 evaluations, meaning that it is absolutely not optimized for the kind of data used in our experiments. This is particularly motivating in this research, since it reflects the complexity of speaker recognition applications such as forensics, dealing with mismatching of development and application data.

3.3. Results for the DAPS evaluation

Figure 1 displays the histograms of the pooled target scores obtained for the DAPS benchmark before and after bias compensation using the model described above. We note that the observed (biased) score distribution is clearly multi-modal. On the other hand, the estimated desired (unbiased) score

distribution resembles a Gaussian shape.

In addition to score compensation, we investigated the effects of classical score normalization on both biased and compensated scores using t-norm [12]. (Although symmetrical scoring is being used, model i-vectors are clean recordings in these experiments and ignored for score normalization purposes.) The reference population for score normalization was obtained from a pool of about 200 tabletop- and body-microphone recordings from the NIST 2002 multi-modal evaluation development data [13]. Figure 2 depicts the distributions of the t-normed target scores before and after compensation. A visual inspection of Figures 1 and 2 suggests that, comparing to bias compensation, score normalization has a secondary effect on reducing the multi-modality caused by trial mismatch.

Next, we will show that, essentially, the impact of score compensation/normalization schemes was on decreasing the score variability across the different conditions, rather than individual performance improvement. Table 1 shows performance in terms of Equal Error Rate (EER) for all the conditions and setups. (A similar trend is observed concerning NIST's Detection Cost Function (DCF) and therefore not presented.) In addition, performance figures for the pooled trials (using a single threshold for all the conditions) in terms of EER and NIST's (old) minimal DCF are listed in Table 2 for the different setups described. By contrasting Tables 1 and 2, we observe that although the performances for individual channels (Table 1) were barely affected by the compensation/normalization setups, significant improvement was achieved in stabilizing the score variability across the pooled conditions (Table 2). A direct implication of this finding is in reducing the dependence on elaborating precise reference score distributions for LLR estimation and calibration.

A further confirmation of the positive impact of the proposed compensation scheme on score stability is suggested by the *Cllr* cost function [14], as follows. We investigated a prospective situation in which scores obtained within a certain condition are calibrated using scores derived from other conditions. Particularly, for each condition, we calculated *Cllr* values based on calibration parameters obtained from each of the other conditions. We estimated the overall mean and standard deviation of these *Cllr* values across all conditions. In fact, since the conditions involving the *balcony* background display relative low detection performance with respect to the other conditions, the above *Cllr* statistics were calculated for two partitions, either including the balcony conditions ("High-noise") or excluding these conditions ("Low-noise").

This experiment was performed for the t-normed biased and unbiased setups and results are shown in Table 3. It can be seen that for the t-norm unbiased setup, the compensated *Cllrs* are consistently lower and more stable across different calibration conditions; and especially concerning the high-noise scenario, for which calibration data is poorly matched.

3.4. Results for the NIST evaluation

In this subsection, we assess the proposed compensation model on the NIST microphone database. In fact, a naive application of our compensation scheme on this database actually led to a slight degradation in accuracy. This degradation can be explained by poor SNR assessment required for estimating the amount of compensation to be applied in each trial, as following.

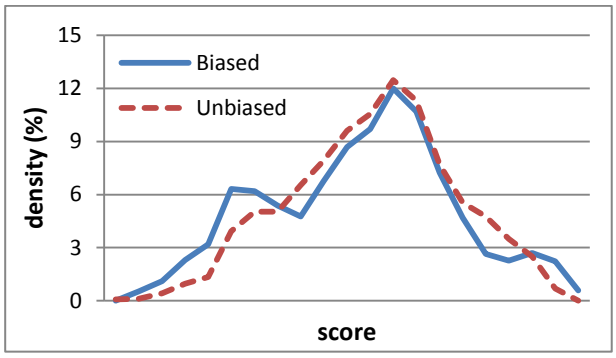

Figure 1: *Histograms of the pooled target scores before and after bias compensation.*

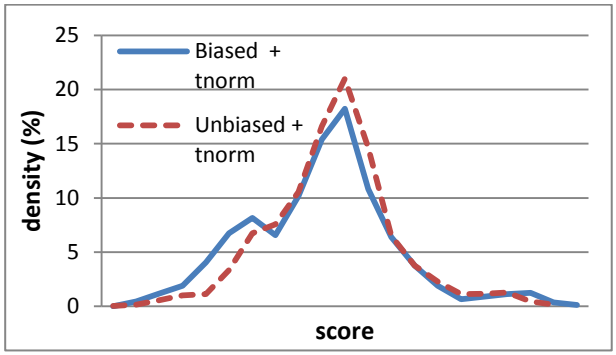

Figure 2: *Histograms of the pooled t-normed target scores before and after bias compensation.*

Table 1. *EERs for biased ("bias"), t-normed-biased ("b+tnorm"), unbiased ("unbias") and t-normed-unbiased ("ub+tnorm") setups across all conditions.*

Condition	EER (%)			
	bias	**b+tnorm**	**unbias**	**ub+tnorm**
ipad_balcony1	14.4	15.5	15.3	16.1
ipad_bedroom1	0.6	1.0	0.5	1.4
ipad_confroom1	0.5	1.0	0.5	1.0
ipad_confroom2	0	0	0	0
ipad_livingroom1	0.1	0.9	0.4	0.9
ipad_office1	2.9	2.0	2.9	2.0
ipad_office2	4.5	5.4	4.4	5.5
ipadflat_confroom1	1.1	0.9	1.0	1.0
ipadflat_office1	1.7	1.0	1.9	1.1
iphone_balcony1	18.1	20.5	16.8	18.0
iphone_bedroom1	1.5	2.1	1.3	2.0
iphone_livingroom1	4.5	4.0	4.5	4.0
Produced	0	0.5	0	0.5

Table 2. *Pooled trials performance for distinct setups.*

Setup	EER (%)	DCF (x10^4)
Biased	16.0	438
Biased+tnorm	14.0	419
Unbiased	10.7	375
Unbiased+tnorm	9.1	355

Table 3. *Biased and unbiased Cllr statistics for low- and high- noise settings*

Setup	*Cllr* mean ± std (Low-noise)	*Cllr* mean ± std (High-noise)
Biased+tnorm	0.60 ± 0.93	1.24 ± 2.07
Unbiased+tnorm	0.52 ± 0.84	0.70 ± 0.97

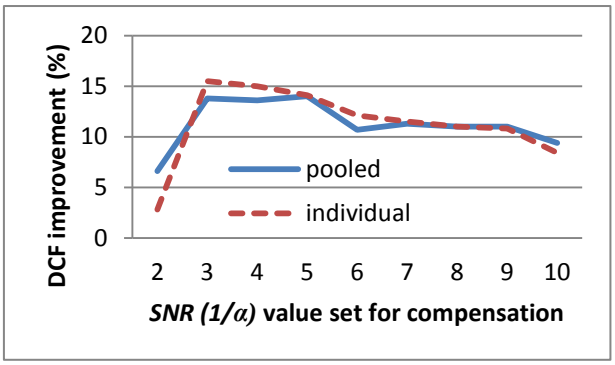

Figure 3. *DCF improvement using fixed SNRs estimates (1/α) for the individual microphones and for the pooled trials.*

Initially, the current VAD actually introduced a certain amount of misclassified speech/non-speech excerpts in this experiment. Secondly, SNR ranges for different microphones were observed to be highly overlapping, as opposed to SNR values obtained for the DAPS database, which were observed to be idiosyncratic to different scenarios. Finally, our model does not account for uncompensated convolutive noise which is significant for the NIST dataset. Therefore, for the moment, lacking a more appropriate noise assessment strategy addressing additive and convolutive effects, we avoid SNR estimation, using pre-fixed SNR values in our compensation scheme for all NIST trials. Score normalization using t-norm as described also led to a slight decrease in performance in this experiment. A more suitable reference population should be selected and therefore normalization and calibration results for this database are not reported.

Figure 3 depicts the relative DCF improvement attained through score compensation for different SNR settings (1/α). (EER gains were marginal.) We show both the average gains across the individual microphone evaluations and the gains for the pooled trials. It can be seen that both curves show similar behavior, contrary to the DAPS experiment, where compensation benefits were obtained for the pooled trials performance rather than for the individual evaluations. This can be explained by the use of a fixed SNR which precludes differential bias compensation across different channels. Nevertheless, since the noise vector incidentally captures channel effects, general performance gains can still be attained for moderate SNR presets.

4. Vector-space SNR

The quality of speech recordings, being one of the major sources of bias in the recognition scores, can be used to identify unsuitable trials [15-18]. Reliable quality measures play an important role in applications such as forensics where potential bias is critical and must be identified in advance.

Table 4. *Correlation between trial quality and recognition score, using distinct quality formulations*

Trial quality measurement	{quality x score} correlation	
	Target trials	Impostor trials
$(1-\alpha_1)(1-\alpha_2)$	0.67	0.36
$n_1 \cdot n_2$	0.76	0.46
$(1-\alpha_1)(1-\alpha_2)n_1 \cdot n_2$	0.80	0.48

As opposed to traditional SNR, in the context of this research, we propose a vectorial SNR quality assessment for a trial, expressed by the dot product between the model and test noise vectors, $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$:

$$SNR_v = n_1 \cdot n_2 \tag{9}$$

The motivation for the vector-based SNR assessment is that it should reflect more accurately the ultimate noise impact on the i-vector space and in particular on the recognition score. Given a speaker recognition trial involving similar noise patterns, their corresponding noise i-vectors, being high correlated, will positively bias the recognition score, as discussed in Section 2.

In order to illustrate this point, we calculated the correlation between the uncompensated pooled scores and the corresponding trial quality measures. Trial quality was expressed through different formulations involving the traditional and vectorial SNR measurements. In particular, we used the inverse of $\boldsymbol{c_1}$ in (7) which reflects the combined SNR of the trial; $\boldsymbol{SNR_v}$, defined in (9); and, in addition, the product of these two measurements. Table 4 summarizes the correlation results for the distinct quality measurements for target and impostor trials in the DAPS evaluation. These correlation levels express a relatively strong relation between the test SNR (due to different test conditions) and the raw recognition score. In these experiments, the proposed vectorial SNR measure outperformed the standard SNR as a quality measure for the speech samples.

5. Conclusions and Future Research

We presented a simple model for compensating noise score bias in the i-vector space for speaker recognition. The methodology relies on the direct parameterization of background noise excerpts in the i-vector space. We showed that noise can be successfully modeled as an interfering vector component in the i-vector space. The estimated noise vector can then be used to compensate the recognition score. Our score compensation scheme weights the amount of noise reduction to be applied according to the estimated SNR on the trial signals. Although the model was originally designed to tackle additive noise, our experiments suggested that channel noise could be indirectly mitigated as well.

In addition, we evaluated the usage of noise i-vectors in a more reliable quality measure for recognition trials, in comparison with regular SNR assessment. Our results showed a better correlation between raw recognition scores and the vectorial noise measurements, in contrast with regular SNR.

Encouraging results show that our method is able to significantly decrease the score variability across different conditions, which is very important for LLR calculation and calibration. Our current results are based on cosine similarity between i-vectors. In fact, other state-of-the-art scoring

techniques such as PLDA should be also investigated.

In part, our experiments emphasized the need of carefully selecting suitable reference populations for score normalization and calibration in mismatched conditions, as typically found in forensic applications [19]. In this regard, note that our method, not being data-driven, may be well suited for uncontrolled scenarios.

Two distinct databases comprising various noise sources were used in the experiments. The databases replicate a variety of scenarios emulating human-machine interaction. Nevertheless, extensive evaluation should be done in order to comprehensively assess the methodology and the various theoretical assumptions involved. In particular, convolutional noise assessment should be formally integrated into this framework.

6. References

[1] 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.

[2] K. S. Rao, S. Sarkar, "Robust speaker recognition in noisy environments", Springer, 2014.

[3] H. Aronowitz, D. Irony, D. Burshtein, "Modeling Intra-Speaker Variability for Speaker Recognition", in Proc. *Interspeech*, 2005.

[4] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition", in *IEEE Transactions on Audio, Speech and Language Processing* 15 (4), pp. 1435-1447, May 2007.

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

[6] W. Kheder, D. Matrouf, J. F. Bonastre, M. Ajili, P. M. Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in Proc. *ICASSP*, 2015.

[7] A. Acero. "Acoustical and Environmental Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1993.

[8] J.-C. Junqua and G. van Noord (eds.), "Robustness in language and speech technology", Kluwer Academic Publishers, 2001.

[9] Description of the DAPS corpus. Available online: https://archive.org/details/daps_dataset

[10] Description of the NIST 2005 SRE. Available online: www.nist.gov/speech/tests/spk/2005

[11] P. Schwarz, P. Matejka, J. Cernocky, "Hierarchical Structures of Neural Networks for Phoneme Recognition", in Proc. *ICASSP*, 2006.

[12] R. Auckenthaler, M., Carey, H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems", *Digital Signal Process*., 10, 42–54, 2001.

[13] Description of the NIST 2002 SRE. Available online: http://www.itl.nist.gov/iad/mig/tests/spk/2002/.

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

[15] Y. Solewicz and M. Koppel, "Considering speech quality in speaker verification fusion," in Proc. *Interspeech*, Lisbon, Sept. 2005.

[16] J. Richiardi, A. Drygajlo, P. Prodanov "Speaker verification with confidence and reliability measures", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, IEEE, Toulouse, France, pp. 641–644, 2006.

[17] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, J. Ortega-García, "Using quality measures for multilevel speaker recognition", Computer Speech and Language, 20 (2-3), pp. 192–209, 2006.

[18] J. Villalba, A. Ortega , A. Miguel , E Lleida, "Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions", Speech Communication Volume 78, 2016.

[19] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition: A need for caution," IEEE Signal Processing Mag., vol. 26, no. 2, pp. 95–103, 2009.