Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Text Dependent Speaker Verification Using a Small Development Set

Hagai Aronowitz

IBM Research - Haifa
Haifa, Israel
hagaia@il.ibm.com

Abstract

Voice biometrics for user authentication is a task in which the object is to perform convenient, robust and secure authentication of speakers. Recently we have investigated the use of state-of-the-art text-independent and text-dependent speaker verification technology for user authentication and obtained satisfactory results within a framework of a proof of technology. However, the use we have made of a quite large development set limits the practical potential of our system. In this work we investigate the ability to build an accurate user authentication system with the limitation of having a small development set.

1. Introduction

With the rapid growth of mobile internet and smart phones, security shortcomings of mobile software and mobile data communication have shifted the focus to strong authentication. Recent advances in voice biometrics offer great potential for strong authentication in mobile environments using voice. This is of particular interest in the financial and banking industry, where financial institutes are looking for ways to offer mobile customers flexible and easy authentication while maintaining security and significantly reducing fraudulent usage.

Recently, a work [1] has been done at IBM within the framework of a proof of technology (POT) which was performed on data collected by the Wells Fargo (WF) bank. The focus of the POT was mainly the evaluation of three text-dependent authentication scenarios. For the best authentication scenario an Equal Error Rate (EER) of 0.6 was obtained (using a global 10-digit string) for the channel matched condition.

However, for many other potential customer engagements the WF POT setup is unrealistic. The development dataset used in the WF POT consisted of 200 recorded speakers with 4 sessions per speaker. A more realistic development set was therefore specified which consists of publicly available text independent (NIST) development data and a reduced text dependent development set consisting of 100 speakers from the WF-POT corpus with only a single session per speaker.

In this paper we present our efforts for building a state-of-the-art text dependent system using the reduced development set specified above.

The remainder of this paper is organized as follows: Section 2 describes the datasets. Section 3 describes our speaker verification systems. Section 4 describes our approach for building our speaker verification systems using a reduced development set. Section 5 presents the results for the individual and fused systems. Finally, Section 6 concludes.

2. Datasets

2.1. Authentication conditions

In the context of text dependent user authentication we defined three different authentication conditions. In the first authentication condition named *global*, a common text is used for both enrollment and verification. In the second condition named *speaker* a user (speaker) dependent password is used for both enrollment and verification. The third condition named *prompted* is a condition in which during the verification stage the user is instructed to speak a prompted text. Enrollment for the *prompted* condition uses speech corresponding to text different than the prompted verification text.

The global condition has the advantage of potentially having development data with the same common text. The speaker condition has the advantage of high rejection rates for imposters who do not know the password. However, in our experiments we assume that the imposters do know the passwords. The prompted condition has the advantage of robustness to recorded speech attacks compared to the global and speaker conditions.

For a proof of technology the WF bank collected data from 750 of its employees. For the *global* condition the WF dataset consists of several common texts. In this work we report results on a single common 10 digit string. For the *speaker* condition, the dataset consists of four speaker dependent passwords, each one used by a quarter of the speakers. However, in order to focus on the scenario of a knowledgeable impostor, we report results for four globally spoken texts which are 10 digit strings. The difference between our *global* condition experiments and our *speaker* condition experiments (besides the different choice of digit strings) is that for the *speaker* condition we assume that development data which contains the chosen digit strings is unavailable. For the *prompted* condition the WF dataset contains an 8-digit string for verification.

2.2. The WF corpus

The WF corpus consists of 750 speakers which are then partitioned into a development dataset (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. Table 1 describes the datasets used for the different conditions. Each session consists of 3 repetitions for each *global* password and 3 repetitions for each *speaker* password. We use all 3 repetitions for *global* and *speaker* enrollment, and only a single repetition for verification for all

authentication scenarios. In all of our experiments we use only same gender trials though the identity of the gender is not assumed to be known by the system. We denote the WF-POT development set by $\mathbf{WF_F}$.

Table 1. Lists of the spoken items used for development, enrollment and verification by the different authentication conditions in the WF evaluation. n_1-n_9 denote 9 distinct 10-digit phone numbers.

Condition	Development spoken items	Enroll spoken items	Eval spoken items
Global	0123456789		
Speaker 1st subset	0123456789 n_1-n_5	n_6	
Speaker 2nd subset		n_7	
Speaker 3rd subset		n_8	
Speaker 4th subset		n_9	
Prompted	n_1-n_6	0123456789 n_1, n_4	25703580

2.3. Standard telephony development set

We use the following standard conversational telephony datasets: Switchboard-II, NIST 2004, 2005 and 2006 speaker recognition evaluations (SREs). We denote this development set by NIST.

2.4. Reduced development dataset

A subset of the WF-POT development set was created by selecting 100 speakers randomly. For 50 of these speakers a single landline session was selected for each speaker. For the other 50 speakers, a single cellular-phone session was selected. In total, the reduced development set consists of 100 sessions. We denote the reduced development set by $\mathrm{WF_R}$.

3. Speaker verification systems

In this section we describe the four speaker verification systems we use in conjunction, and our fused system.

3.1. JFA-based system

Our Joint Factor Analysis (JFA)-based system is inspired by the theory described thoroughly in [2]. A detailed description of our implementation can be found in [3]. Differently from the standard implementation, we use the following two variants to better cope with short and asymmetric sessions (enrollment longer than test).

First we use a robust scoring function (Equation 1) which gives an average relative error reduction of 8% for our text dependent scenarios.

$$LLR_{robust} = \frac{s_E{}^t N_E{}^{\frac{1}{2}} N_T{}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-1} s_T}{tr(N_E{}^{\frac{1}{2}} N_T{}^{\frac{1}{2}})} \quad (1)$$

In Equation 1 s_E denotes the centered and compensated supervector for the enrollment session $s_E = Vy_E + Dz_E$ and s_T denotes the centered compensated supervector for the test session $s_T = N_T^{-1} F_T - Ux_T - m$. V, D and U stand for the speaker, common and channel JFA hyper-parametric matrices, y_E and z_E are point estimates for the speaker and common factors for the enrollment session, x_T is a point estimate for the channel factors for the test session, F_T is a vector consisting of the first order statistics for the test session, and N_E and N_T are the zero order statistics for the enrollment and test sessions correspondingly, arranged in matrices as explained in [3]. Finally, m stands for the UBM (Universal Background Model) supervector, and $\sum$ is a block matrix with covariance matrices from the UBM on the diagonal.

Our second deviation from standard JFA is the use of an asymmetric combination of forward and reverse scores using a simple weighting scheme. The weighed fusion method enables us to gain from reverse scoring even when test sessions are shorter than the enrollment session (the WF POT typical scenario).

Our JFA-based system was built using 12,711 sessions from Switchboard-II, NIST 2004 SRE and NIST 2006 SRE. The reason we did not use the WF POT development data is that when doing that, we observed only a small improvement compared to using the standard conversational telephony data. The only use we made of the WF POT development data is for ZT-score normalization.

3.2. I-vector-based system

Our i-vector-based system [4] is inspired by the work described in [5]. We use standard i-vector extraction with length normalization followed by LDA (Linear Discriminant Analysis) and WCCN (Within Class Covariance Normalization). We use cosine-based similarity scoring and normalize using ZT-norm which we found to be slightly superior to s-norm in our setup. The development data used for system building is identical to the data we use for JFA building.

3.3. GMM-NAP-based system

Our GMM-NAP system inspired by [6] is described in detail in [1]. Our GMM-NAP system deviates from the standard by the following modifications.

3.3.1. Two-wire NAP

In [7, 8] we discovered that removing dominant components of the inter-speaker variability subspace in addition to removing the intra-speaker inter-session variability subspace improves speaker recognition accuracy not only for 2-wire data (for which this method was originally designed) but also for regular 4-wire data. This variant named 2-wire-NAP is therefore part of our baseline GMM-NAP system and led to a relative reduction of 6% in EER on the WF POT.

3.3.2. Text dependent UBM & NAP projection

Contrary to the JFA and i-vector frameworks, NAP requires smaller quantities of development data to properly estimate the hyper-parameters (UBM and NAP projection). In [1] it was found that estimating text-dependent UBM and NAP from the

WF-POT development set led to a relative reduction of 37% in EER.

3.3.3. Geometric mean comparison kernel

Contrary to [1], we now use the kernel introduced in [9] for scoring a pair of sessions:

$$C_{GM}(E,T) = m_E^t\left(\lambda_E^{1/2} \otimes I_n\right)\Sigma^{-1}\left(\lambda_T^{1/2} \otimes I_n\right)m_T \qquad (2)$$

where E and T stand for the enrollment and test sessions, m_E and m_T are the corresponding concatenated GMM means, λ_E and λ_T are the corresponding concatenated GMM weights, $\sum$ is a block matrix with covariance matrices from the UBM on the diagonal, n is the feature vector dimension, and $\otimes$ is the Kronecker product.

3.4. HMM-NAP-based system

The HMM-NAP-based system is an extension of the GMM-NAP system in the sense that instead of using a UBM to parameterize audio sessions into GMM-supervectors, a speaker-independent (SI) Hidden Markov Model (HMM) is used to parameterize audio sessions into HMM-supervectors. The other components of the GMM-NAP system (feature extraction, 2-wire-NAP estimation and compensation, dot-product scoring and ZT-normalization) are used identically in the HMM-NAP framework.

We use our HMM-NAP system for the *global* authentication condition (shared password) only. For a given shared password a SI-HMM is trained using all repetitions of the shared password in the development data. The SI-HMM is then used to parameterize all the repetitions of the shared password in the development, train and test datasets. We use only the Gaussian means of the different HMM states (with a similar normalization as done for the GMM-NAP system) for supervector creation.

3.5. Fused system

We combine the scores of the JFA, i-vector, GMM-NAP and HMM-NAP (for the global condition) into a single fused system. The scores are combined using a weighted average which assigns a double weight for systems which are significantly more accurate.

4. Speaker verification using a reduced development dataset

Our JFA and i-vector based systems use the WF-POT development data for score normalization only. The effect of the reduction in text dependent development data is therefore limited to a modest degradation in accuracy due to less accurate score normalization.

Our NAP-based systems are affected significantly by the reduction in text dependent development data. Not only that the amount of data used to train the UBM (or SI-HMM), estimate the NAP projection and estimate the ZT-norm statistics is reduced, but more severely the ability to capture intersession variability from the development data is significantly reduced because intersession variability can no longer be isolated from other variabilities such a speaker variability.

In the following subsections we report our efforts for building our NAP-based system with the reduced text dependent development data.

4.1. Building the GMM-NAP system with the reduced development dataset WF_R

4.1.1. UBM training

We investigate two options. The first one is training the UBM using NIST which is large enough but does not match the evaluation text. The second option is training the UBM using WF_R.

4.1.2. NAP estimation

WF_R is inappropriate for estimation a standard NAP projection because it does not contain multi-session speakers. Our first option is to estimate the NAP projection from NIST.

The second option is to estimate the NAP projection from WF_R by using the *common speaker subspace* (*CSS*) compensation method introduced in [7]. According to this method, a *CSS* is estimated from a large set of sessions using kernel-PCA. The *CSS* is removed from the sessions thus producing sessions located in the *speaker unique subspace (SUS)*. The *SUS* is supposed to consist of information that is not common to many speaker (otherwise it would be captured by the *CSS*) and therefore should be appropriate for speaker discriminating. According to [7] *CSS* removal was almost as good as using standard NAP compensation. Using the GMM supervector framework, the *CSS* removal method is equivalent to NAP compensation with the NAP projection estimated by applying PCA analysis to the pooled supervectors from the entire development set (without use of speaker label).

A third option is to use both the NAP projection estimated from NIST and *CSS*-removal estimated from WF_R to compensate both subspaces.

4.2. Building the HMM-NAP system with the reduced development dataset WF_R

4.2.1. SI-HMM training

The SI-HMM is text dependent. Therefore, we only train the SI-HMM using WF_R.

4.2.2. NAP estimation

Contrary to the GMM-NAP framework, the NAP projection is text-dependent. Therefore, the only option is to estimate it from WF_R using the *CSS* compensation method.

4.2.3. fNAP

In order to make use of the text independent development set (NIST), the fNAP [10] method is used for channel compensation in the feature domain. This can be either done exclusively or in conjunction with NAP compensation (estimated from WF_R).

5. Results

5.1. JFA & i-vector based results

Tables 2 and 3 present a comparison of the JFA and i-vector based systems using either the full or the reduced development sets. The results for the channel matched condition are reported in Table 2, and the results for the channel mismatched condition are reported in Table 3. The relative EER increase due to the reduction of the development set used to estimate ZT statistics (from 800 sessions to 100 sessions) is 20% in average for the matched condition and 13% for the mismatched condition.

Table 2. A comparison of the JFA and i-vector systems using the full (in bold) and reduced development sets. Target trials are channel matched. Results are in EER.

System	*Global*	*Speaker*	*Prompted*
JFA WF_F	**1.25**	**1.76**	**5.13**
JFA WF_R	1.55	2.13	6.33
i-vector WF_F	**1.69**	**2.19**	**5.44**
i-vector WF_R	2.03	2.81	5.89

Table 3. A comparison of the JFA and i-vector systems using the full (in bold) and reduced development sets. Target trials are channel mismatched. Results are in EER.

System	*Global*	*Speaker*	*Prompted*
JFA WF_F	**3.57**	**4.48**	**10.99**
JFA WF_R	4.06	5.17	11.72
i-vector WF_F	**4.71**	**5.78**	**11.06**
i-vector WF_R	5.41	6.82	11.85

5.2. GMM-NAP-based results

Tables 4 and 5 present a comparison of the different GMM-NAP system builds described in Section 4. The results for the channel matched condition are reported in Table 4, and the results for the channel mismatched condition are reported in Table 5.

Table 6 presents a comparison of the mean relative EER degradation for the different system builds. In general, training both the UBM and the NAP projection on the NIST data was found to be the worst policy and resulted in an average relative increase of 37% in EER, compared to using the whole WF development dataset.

Training the UBM on the WF reduced development set and estimating the NAP projection from the NIST development set was found to be better (especially for the mismatched condition), and so was training both the UBM and the NAP projection on the WF reduced dataset (using *CSS* removal) (especially for the matched condition). The combined method of training a UBM on the WF reduced dataset and applying both a NAP projection estimated on NIST data and *CSS* removal estimated from the WF reduced dataset was found to be best.

Using the combined method, the following results were obtained. For the *global* condition the relative EER increase due to the reduction of the development set is around 40% for both channel matched and mismatched conditions. For the *speaker* and *prompted* conditions the relative EER increase due to the reduction of the development set is around 5% for the channel matched condition, and around 20% for the channel mismatched condition. The reason we get a larger degradation for the *global* condition is that the *global* condition benefits mostly from text matching development data.

Table 4. EERs for the proposed GMM-NAP methods using the full (in bold) and reduced development sets. Target trials are channel matched.

GMM-NAP system			*Global*	*Speaker*	*Prompted*
UBM	NAP	ZT			
WF_F			**0.83**	**1.54**	**4.39**
NIST		WF_R	1.83	3.11	5.95
WF_R	NIST	WF_R	1.28	1.88	4.65
	WF_R		1.24	1.78	4.44
	WF_R+ NIST		1.17	1.65	4.55

Table 5. EERs for the proposed GMM-NAP methods using the full (in bold) and reduced development sets. Target trials are channel mismatched.

GMM-NAP system			*Global*	*Speaker*	*Prompted*
UBM	NAP	ZT			
WF_F			**2.33**	**4.15**	**9.22**
NIST		WF_R	4.39	7.49	12.22
WF_R	NIST	WF_R	3.66	5.40	11.47
	WF_R		4.33	5.74	12.94
	WF_R+ NIST		3.40	5.04	11.32

Table 6. A list of the mean degradation for the proposed GMM-NAP methods compared to using the full (in bold) development set.

GMM-NAP system			*Mean EER rel. increase (in %)*	
UBM	NAP	ZT	Matched	Mismatched
WF_F			-	
NIST		WF_R	38	35
WF_R	NIST	WF_R	13	24
	WF_R		9	32
	WF_R+ NIST		8	21

5.3. HMM-NAP-based results

Table 7 presents a comparison of the different HMM-NAP system builds described in Section 4. It can be seen that estimating the NAP projection on the WF reduced dataset (using *CSS* removal) outperformed using fNAP estimated on the NIST development data. The combination of both was found to be best and results in a relative increase of 33% and 66% in EER for the channel matched and channel mismatched conditions respectively.

In order to better understand the sources for degradation, we ran some more experiments trying to isolate the degradation due to the use of a reduced development set for the SI-HMM training, NAP estimation and ZT-score

normalization. The results are reported in Table 8. We can see that most of the degradation is due to the inappropriate estimation of the NAP projection.

Table 7. EERs for the proposed HMM-NAP-based methods for the global condition using the full (in bold) and reduced development sets.

HMM-NAP system			Matched channel	Mismatched channel
SI-HMM	NAP	ZT		
WF_F			**0.84**	**1.98**
WF_R	WF_R	WF_R	1.17	3.46
	fNAP		1.35	3.74
	WF_R+ fNAP		1.12	3.28

Table 8. An analysis of the sources of degradation for our proposed HMM-NAP-based method for the global condition using the reduced development set. Results are in EER.

HMM-NAP system			Matched channel	Mismatched channel
SI-HMM	NAP	ZT		
	WF_F		**0.84**	**1.98**
WF_R		WF_R	0.90	2.32
	WF_R+ fNAP	**WF_F**	1.09	3.18
		WF_R	1.12	3.28

5.4. Fused system results

Tables 9 and 10 present a comparison of the fused system using the full and reduced development set. The average relative EER increase due to the use of a reduced development set is 20%.

Table 9. A comparison of the fused system using the full and reduced development sets. Target trials are channel matched. Results are in EER.

Condition	Fused full dev. set	Fused reduced dev. Set	Relative degradation (in %)
Global	0.56	0.66	18
Speaker	0.85	0.97	14
Prompted	2.48	3.05	23

Table 10. A comparison of the fused system using the full and reduced development sets. Target trials are channel mismatched. Results are in EER.

Condition	Fused full dev. set	Fused reduced dev. set	Relative degradation (in %)
Global	1.56	1.97	26
Speaker	2.87	3.24	13
Prompted	6.41	7.87	23

6. Conclusions

In this work we have explored the possibility of building our text dependent speaker verification systems on a small text dependent development set consisting of only 100 sessions (from 100 distinct speakers). We have managed to obtain a relatively small degradation in accuracy (20% relative increase in EER) compared to when using the full development set (a total of 800 sessions from 200 speakers).

Our JFA and i-vector based systems do not make a strong use of text dependent development data anyway (except for score normalization). We intend to change that in the future.

Our NAP-based systems are totally dependent on text-dependent development data. Regarding the UBM, SI-HMM and score normalization, the reduced developments set was found to be sufficient (though not ideal) because they lack of a need for a multiplicity of sessions per speaker in the development set.

NAP estimation is inherently dependent on the availability of a multiplicity of sessions per speaker in the development set. We therefore replace the conventional NAP estimation method with a combination of the following two compensation methods. First, NAP estimated from a standard text independent dataset (NIST) is applied directly by the GMM-NAP system or indirectly (using fNAP [10]) by the HMM-NAP system. Second, *common speaker subspace* removal (CSS) [7] estimated from the small text dependent development set is applied by both the GMM-NAP and HMM-NAP system.

Efficient use of available text-dependent development data proves to be important for obtaining high accuracy in text dependent speaker verification. We plan to improve our use of such data, especially for the JFA and i-vector based systems.

7. Acknowledgements

The author wishes to thank Wells Fargo for collecting and providing the data for the feasibility study.

8. References

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

[2] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms", technical report CRIM-06/08-14, 2006.

[3] H. Aronowitz, O. Barkan, "New Developments in Joint Factor Analysis for Speaker Recognition", in Proc. *Interspeech*, 2011.

[4] H. Aronowitz. O. Barkan, "Efficient Approximated I-Vector Extraction", to appear in *ICASSP*, 2012.

[5] 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.

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

[7] H. Aronowitz, "Speaker Recognition using Kernel-PCA and Intersession Variability Modeling", in Proc. *Interspeech*, 2007.

[8] Y. A. Solewicz, H. Aronowitz, "Two-Wire Nuisance Attribute Projection", in Proc. *Interspeech*, 2009.

[9] W. Campbell, Z. Karam, "Simple and Efficient Speaker Comparison using Approximate KL Divergence", in Proc. *Interspeech*, 2010.

[10] W. M. Campbell, D. E. Sturim, P. A. Torres-Carrasquillo, D. A. Reynolds, "A Comparison of Subspace Feature-Domain Methods for Language Recognition", in Proc. Interspeech, 2008.