ISCA Archive
http://www.isca-speech.org/archive

Afeka-AVIOS
Speech Processing Conference 2012
Tel Aviv, Israel
June 19-20, 2012

Voice Biometrics for User Authentication

Hagai Aronowitz

IBM Research – Haifa, Haifa, Israel

hagaia@il.ibm.com

Abstract

Voice biometrics for user authentication is a task in which the goal is to perform convenient, robust and secure authentication of speakers. In this work we investigate the use of state-of-the-art text-independent and text-dependent speaker verification technology for user authentication. We evaluate three different authentication conditions: global digit strings, speaker specific digit stings and prompted digit strings. Harnessing the characteristics of the different types of conditions can provide benefits such as authentication transparent to the user (convenience), spoofing robustness (security) and improved accuracy (reliability). The systems were evaluated on a corpus collected by Wells Fargo Bank which consists of 750 speakers. We show how to adapt techniques such as joint factor analysis (JFA), i-vectors, Gaussian mixture models with nuisance attribute projection (GMM-NAP) and hidden Markov models with NAP (HMM-NAP) to obtain improved results for new authentication scenarios and environments.

Overall, EERs significantly lower than 1% have been obtained for the matched channel condition, while the error almost triples for the mismatched channel condition.

In order to be able to use advanced techniques such as JFA and i-vectors in a realistic low-latency system we have developed the JFAlight method and the efficient i-vector extraction method for efficient approximated JFA and i-vector scoring. Using these algorithms we managed to speed up the JFA and i-vector methods to be comparable to the widely used NAP method.

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. The existing user-id/password methodology, while tolerable for desktops and laptops, is inadequate for mobile use due to the difficulty of data entry on a small form factor device and a higher risk of the device getting in the hands of unauthorized users. 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.

This paper describes the work done at IBM within the framework of a proof of technology (POT) which was performed on data collected by Wells Fargo. Although most of the evaluated authentication scenarios are text-dependent we mostly used text-independent speaker verification technology (namely JFA [1], i-vectors [2] and GMM-NAP [3]) for the POT. The only exception made was in the case of user authentication using a fixed common digit-string where we used text-dependent speaker verification technology (namely HMM-NAP [4]) in conjunction with the text-independent technology. However, in order to benefit from the particular characteristics of the data we adapted the GMM-NAP-based system and to a lesser extent also the JFA-based and i-vector systems to the development data we were provided within the POT framework.

The remainder of this paper is organized as follows: Section 2 describes the datasets. Section 3 describes our JFA, i-vector and GMM-NAP-based text-independent systems and our HMM supervector-based text-dependent system. Section 4 presents the results for our individual and fused systems. Section 5 reports accuracy and speed measurements using approximated JFA and approximated i-vector extraction. 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.

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 12,711 sessions from the following datasets: Switchboard-II, NIST 2004, 2005 and 2006 speaker recognition evaluations (SREs).

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 [1]. A detailed description of our implementation can be found in [5]. 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 [5]. 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 the telephony development set described in subsection 2.3. 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 [6] is inspired by the work described in [2]. 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 [3] is described in detail in [4]. 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). As we reported in [4], estimating text-dependent UBM and NAP from the WF-POT development set led to a dramatic reduction in EER.

3.3.3. Geometric mean comparison kernel

Contrary to [4], 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. Results

In this section we report the results for the three authentication conditions using JFA, i-vector, GMM-NAP, HMM-NAP (for *global* only) and the fused system which is obtained by taking an average of the JFA, i-vector and GMM-NAP scores (for the global condition, the fused JFA, i-vector and GMM-NAP score is further averaged with the HMM-NAP score).

Tables 2 and 3 present results for the channel matched and channel mismatched conditions respectively. For all conditions the NAP-based systems outperform both the JFA and the i-vector systems due to the fact that the GMM-NAP was built on the WF-POT development dataset and the JFA and i-vector systems were mostly built on conversational telephony.

Table 2. *EER (in %) for the three authentication conditions. Target trials are channel matched.*

Condition	JFA	i-vector	GMM NAP	HMM NAP	Fused
Global	1.25	1.69	0.83	0.84	0.56
Speaker	1.76	2.19	1.54	-	0.85
Prompted	5.13	5.44	4.39	-	2.48

Table 3. *EER (in %) for the three authentication conditions. Target trials are channel mismatched*

Condition	JFA	i-vector	GMM NAP	HMM NAP	Fused
Global	3.57	4.71	2.33	1.98	1.56
Speaker	4.48	5.78	4.15	-	2.87
Prompted	10.99	11.06	9.22	-	6.41

5. Approximated JFA and i-vector extraction

In [5] we introduced the JFAlight method which manages to speed up LLR calculation under the JFA framework (excluding sufficient statistics calculation which is relatively fast) by a factor of 100 with no statistically significant degradation in accuracy for the WF evaluation.

In [6] we introduced an efficient method for approximated i-vector extraction. The method manages to speed up i-vector extraction (excluding sufficient statistics calculation which is relatively fast) by a factor of 25 with no statistically significant degradation in accuracy for the WF evaluation.

6. Conclusions

In this work we explored three different user authentication conditions namely *global*, *speaker* and *prompted*. We evaluated four speaker recognition frameworks (JFA, i-vector, GMM-NAP and HMM-NAP) and a fusion of the four. The HMM-NAP algorithm was found to be the best single system for the *global* condition. Our GMM-NAP system which is inferior to our JFA system on a standard NIST SRE (EER=3.6% compared to EER=1.4% on NIST-2008 data) was superior on the WF POT evaluation due to its full usage of the WF POT development data. We managed to improve our baseline GMM-NAP system significantly mostly by using the most appropriate data for UBM and NAP-projection estimation.

Overall, EERs lower than 1% have been obtained for the matched channel condition (for *global* and *speaker*), while the error triples for the mismatched channel condition.

Furthermore, fast JFA scoring [5] and fast i-vector extraction [6] has reduced the time complexity of these scoring systems to be comparable to the time complexity of GMM/HMM-NAP scoring with an insignificant degradation in accuracy compared to the original techniques.

7. Acknowledgements

The authors wish to thank Wells Fargo for collecting and providing the data for the feasibility study, and to thanks Jason Pelecanos, Oren Barkan, Ron Hoory and David Nahamoo for their contributions.

8. References

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

[2] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” in Proc. *Interspeech*, 2009.

[3] W. Campbell, D. Sturim, D. Reynolds, A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation", in Proc. *ICASSP*, 2006.

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

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

[6] H. Aronowitz, O. Barkan, "Efficient approximated i-vector extraction", in Proc. *ICASSP*, 2012.

[7] H. Aronowitz, Y. A. Solewicz, “Speaker Recognition in Two-Wire Test Sessions”, in Proc. *Interspeech* 2008,

[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.