New Developments in Voice Biometrics for User Authentication

Hagai Aronowitz\textsuperscript{1}, Ron Hoory\textsuperscript{1}, Jason Pelecanos\textsuperscript{2}, David Nahamoo\textsuperscript{2}

\textsuperscript{1} IBM Research – Haifa, Haifa, Israel
\textsuperscript{2} IBM T.J. Watson Research Center, Yorktown Heights, NY, U.S.A
\{hagai, hoory\}@il.ibm.com, \{jwpeleca, nahamoo\}@us.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. 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 four different authentication conditions: speaker specific digit strings, global digit strings, prompted digit strings, and text-independent. 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), 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.

Index Terms: Speaker verification, speaker recognition, text dependent, text independent, HMM supervectors

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 and GMM-NAP) 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) 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 system 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 and GMM-NAP-based text-independent systems. Section 4 describes our HMM supervector-based text-dependent system. Section 5 presents the results for our individual and fused systems. Finally, Section 6 concludes and describes post evaluation work.

2. Datasets

2.1. Authentication conditions

In the context of user authentication we defined four 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. Finally, in the text independent condition the user is enrolled by reading a fixed text (shared among all speakers) and verified by saying utterances such as user's full name, user work phone number, user's zip code, etc.

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. Finally, the text-independent condition has the advantage of being more natural especially when the audio is implicitly captured during a conversation between the user and an agent.

For a proof of technology the Wells Fargo (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 two common texts which are 10 digit strings. For the speaker condition, the dataset consists of several speaker dependent passwords. 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. For the text independent condition the WF dataset consists of one read global text for enrollment and speaker-dependent texts for
verification. For each speaker the verification text contains the full name, the zip-code, the work phone number, etc.

### 2.2. Datasets

The WF corpus consists of 750 speakers which are 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. Tables 1 and 2 describe 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 (unless stated otherwise). In all of our experiments we use only same gender trials (results for cross gender trials are much better) 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.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Development spoken items</th>
<th>Enroll spoken items</th>
<th>Eval spoken items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global 1</td>
<td>n1</td>
<td>n1</td>
<td>n1</td>
</tr>
<tr>
<td>Global 2</td>
<td>n2</td>
<td>n2</td>
<td>n2</td>
</tr>
<tr>
<td>Speaker 1</td>
<td>n1,n2,n7-n10</td>
<td>n3</td>
<td>n3</td>
</tr>
<tr>
<td>Speaker 2</td>
<td></td>
<td>n4</td>
<td>n4</td>
</tr>
<tr>
<td>Speaker 3</td>
<td></td>
<td>n5</td>
<td>n5</td>
</tr>
<tr>
<td>Speaker 4</td>
<td></td>
<td>n6</td>
<td>n6</td>
</tr>
<tr>
<td>Prompted</td>
<td>n1,n3,n7-n10</td>
<td>n1,n2,n9</td>
<td>n1,n11,n12</td>
</tr>
<tr>
<td>Text Independent</td>
<td>t1-t4</td>
<td>t1</td>
<td>t5</td>
</tr>
</tbody>
</table>

#### Table 2. A description of the different spoken items used in the WF evaluation.

<table>
<thead>
<tr>
<th>Spoken Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>n1</td>
<td>&quot;0123456789&quot;</td>
</tr>
<tr>
<td>n2-n10</td>
<td>10 digit phone numbers</td>
</tr>
<tr>
<td>n11</td>
<td>&quot;2570&quot;</td>
</tr>
<tr>
<td>n12</td>
<td>&quot;3580&quot;</td>
</tr>
<tr>
<td>t1, t2</td>
<td>read texts (~25 sec each)</td>
</tr>
<tr>
<td>t3</td>
<td>a short global sentence</td>
</tr>
<tr>
<td>t4</td>
<td>a short global sentence</td>
</tr>
<tr>
<td>t5</td>
<td>a combination of short phrases such as full name, work phone number etc. (~17 sec)</td>
</tr>
</tbody>
</table>

### 3. Text independent systems

In this section we describe the two text-independent systems we used for all the authentication conditions.

#### 3.1. JFA-based system

Our JFA-based system is inspired by the theory described thoroughly in [1]. JFA is a framework used to jointly model speaker and intra-speaker inter-session variability. In this model, each session can be represented by its zero and first order statistics under a GMM-UBM framework (in some implementations also the second order statistics are used). The basic assumption is that a speaker- and channel-dependent supervector of GMM means denoted by \( M \) can be decomposed into a sum of two supervectors: a speaker supervector \( s \) and a channel supervector \( c \):

\[
M = s + c \tag{1}
\]

where \( s \) and \( c \) are normally distributed according to Equations (2, 3):

\[
s = m + Vy + Dz \tag{2}
\]

\[
c = Ux. \tag{3}
\]

In Equations 2-3, \( m \) is the supervector corresponding to the universal background model (UBM), \( D \) is a diagonal matrix, \( V \) is a rectangular matrix of low rank, and \( y \) (speaker factors) and \( z \) (common speakers) are independent random vectors having standard normal distributions. \( U \) is a rectangular matrix of low rank, \( x \) (channel or inter-session factors) is a vector distributed with a standard normal distribution.

In this work we trained the UBM and the JFA hyper-parameters \( U, V \) and \( D \) on standard conversational telephony data. We used 12,711 sessions from Switchboard-II, NIST 2004 speaker recognition evaluation (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, for which we use all of the 800 development sessions for both Z-norm and T-norm.

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 delta-delta cepstral coefficients extracted every 10ms using a 32ms window. Feature warping is applied with a 300 frame window after computing the delta-features. We used a GMM order of 1024. We used 300 speaker factors and 100 channel factors. Scoring is performed using the quadratic scoring method described in [2] followed by ZT-score normalization.

#### 3.2. GMM-NAP-based system

Our GMM-NAP system inspired by [12] and based on our previous systems described in [3, 4]. In the GMM-NAP framework a GMM is adapted for each session (training, testing and development) from a UBM using MAP-adaptation. NAP is estimated from a development dataset and is used to compensate intra-speaker inter-session variability (such as channel variability). In [4] it was discovered that removing dominant components of the inter-speaker variability subspace on top of 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. We therefore use this variant named 2-wire-NAP in our experiments.

We use 512-order GMMs. 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. We estimate a 50-dimensional 2-wire NAP projection using the normalized supervectors.

Contrary to the JFA framework, the NAP framework requires smaller quantities of development data to properly estimate the hyper-parameters (the NAP projection). We
therefore investigated the possibility of estimating the UBM and/or the NAP projection from the WF POT development data. The results presented in section 5 show that indeed the accuracy can be significantly improved by estimating both the UBM and the NAP projection from text matching the development data. The WF POT development data is also used for ZT-score normalization for which we use all of the 800 development sessions for both Z-norm and T-norm. 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 13 cepstral coefficients augmented by 13 delta cepstral coefficients extracted every 10ms using a 25ms window. Feature warping is applied with a 300 frame window before computing the delta-features. Scoring is performed using a dot-product between the compensated train and test supervectors followed by ZT-score normalization.

4. HMM-based system

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

HMM-supervectors were proposed in [5] and were used as a method for parameterization of word-conditioned audio segments for SVM classification in the context of text-independent speaker verification (in the scenario that a transcription is available). In [6] the use of a NAP compensation component was proposed to improve the system reported in [5].

In [7] a system based on HMM supervector parameterization followed by SVM classification was proposed for the task of text-dependent speaker verification, but no channel (or inter-session variability) compensation was applied.

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. A 50-order 2-wire NAP projection is estimated in a similar way to the GMM-NAP system.

5. Results

5.1. GMM-NAP-based system tuning

In this subsection we show how to configure the GMM-NAP system for the global condition. We evaluate the GMM-NAP system on an extended global dataset using 8 different shared passwords (both digit strings and short phrases). The process of building the GMM-NAP system can be decomposed into three components: training the UBM, estimating the NAP projection, and selecting the T- and Z-norm sessions. Our baseline system is based on training the UBM and the NAP transforms on a separate string digits-based text-dependent development dataset which was collected in 2003 internally by IBM (IBM-2003). For a given shared password we select Z-

and T-norm sessions from the WF POT development dataset which correspond to the same password.

Table 3 presents results for different variants of the GMM-NAP system. Using 2-wire NAP and estimating the UBM and the NAP-projection on the more closely matching development data results in an error reduction of 25% relative compared to the baseline. We use this approach for our GMM-NAP system throughout the rest of the paper.

5.2. WF POT Evaluation Results

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

Tables 4 and 5 present results for the channel matched and channel mismatched conditions respectively. For the global condition the HMM-NAP system outperformed both the JFA and the GMM-NAP systems. For the speaker and prompted conditions GMM-NAP outperformed JFA due to the fact that the GMM-NAP was built on the WF development dataset and the JFA system was mostly built on conversational telephony. For the text-independent condition which relatively matches the conversational telephony development dataset, the JFA system outperformed the GMM-NAP system.

Tables 6 and 7 present results for the channel matched and channel mismatched conditions respectively for the case of verification sessions containing two consecutive repetitions of the password digit string (global and speaker conditions only). Once more, the HMM-NAP system is found to be superior for the global condition. For the speaker condition, the GMM-NAP system is slightly better than the JFA system.

Table 8 presents results for the text-independent condition as a function of verification session length. For these experiments we use both a cellular and a landline session (two sessions in total) for enrolling each speaker.
Overall, EERs lower than 1% have been obtained for the matched channel condition, while the error doubles for the mismatched channel condition. Multi-condition authentication (fusing several authentication conditions) may lead to even better accuracy. For instance, fusion of the global condition (with two verification repetitions) with the text-independent condition (with two enrollment sessions) results with EER=0.36% for the matched channel condition and EER=0.81% for the mismatched channel condition.

Following our submission for the WF POT evaluation we have reduced the EER for the different authentication conditions by roughly 20% relative. This improvement was achieved mostly by the following two activities. The first one is fusing a fourth system which is an i-vector [8] based system which has accuracy comparable to the JFA system. The second improvement has been achieved by modifying both the NAP and the JFA scoring functions according to the $C_{GM}$ scoring function defined in [9].

Recently we have been exploring other aspects essential to the success of a user authentication system, such as automatic goat detection [10] and fast JFA scoring [11] which has reduced the time complexity of JFA scoring to be comparable to the time complexity of GMM-NAP scoring with an insignificant degradation in accuracy.

### Acknowledgements

The authors wish to thank Wells Fargo for collecting and providing the data for the feasibility study.

### 8. References


