Attacks on Speech Biometric Authentication

K. Inthavisas and D. Lopresti

Department of Computer Science and Engineering,
Lehigh University, Bethlehem, PA 18015

Abstract— *We investigate the security of DTW, VQ and GMM methods that have been used in speaker authentication systems. We present attack models based on adversary knowledge. We start with naive adversaries without knowledge of an authentic speaker and develop them into highly knowledgeable adversaries who know the speaker's information, have the speaker's voice samples, acquire the speaker's template, and know an algorithm of the speaker authentication system. We propose an analysis-synthesis forgery in which the informed adversary can exploit information such as feature vectors from the template and a statistical probability from the voice samples of the target speakers to re-generate a forgery that can be used in remote or on-line authentication. We evaluate our scheme with two speech datasets. The results show that the proposed scheme outperforms other attack models reported in the literature.*

Keywords: Speaker verification, Generative attack, Pattern recognition, Biometrics security

1. Introduction

The traditional approach to access a system has used a password to authenticate users. Unfortunately, history has proved that the user tends to use a password that is easy to guess [9]. Moreover, this approach is vulnerable to a social engineering attack [16]. For these reasons, researchers are interested in using biometrics for a user authentication system. However, the weak security of the biometrics system against forgery attacks is a concern.

Two types of adversaries have to be considered: a human or an algorithm [1]. For both types, the ability of the adversary to compromise the system depends on their knowledge of private information, public information and the motivation [3]. The knowledge of private information includes target biometrics and public information includes auxiliary information such as the construction of the authentication system and templates [1]. The motivation depends on how important gaining access to the system is to the adversary.

In our case, we choose to study speech-based biometrics. Several reasons for investigating the security of speech-based biometrics are: 1) The system is inexpensive compared with the implementation of other kinds of biometrics (e.g., iris or fingerprints). 2) Voice is a behavioral biometric; users can change their pass-phrases easily. 3) Attacks against speech biometric templates have not be studied as much as attacks against other kinds of biometrics (e.g., fingerprints or handwriting).

In this paper, we investigate the security of user authentication based on the knowledge of private and public information and the motivation of the adversary. Three systems based on a pattern matching technique are used in the study. According to [4], the pattern matching methods include a template, a codebook, and a statistical model. Dynamic Time Warping (DTW) is used in the template model. Vector Quantization (VQ) in the codebook model, and Hidden-Markov Model (HMM) in the statistical model. For statistical models, we used a single state HMM which is referred to as a Gaussian Mixture Model (GMM) [4].

We attack the systems using human and algorithmic attacks. For the first scenario (human), we ask a subject to say the pass-phrases of the target users for multiple rounds. For the first round, the impostors say the pass-phrases without listening to the target voice. In the second round, they are asked to imitate the pass-phrases of the target users by listening to the voice of the target users. In this round, we ensure that the subjects are well-motivated by providing an incentive reward for the best imitator. For the second scenario (algorithmic), we will use voice recordings from the target users to generate synthesized pass-phrases. The synthesized sound will be generated from state-of-the-art technologies; we use HMM-based speech synthesizer. We carefully designed the collection of the voice data, so the voice would not overlap with the pass-phrases of the target users. In the last scenario (algorithmic), we re-generate users' pass-phrases based on the template information. Then, these pass-phrases will be used to attack the systems. These scenarios are detailed in section IV-4.2.

2. Related work

Many biometrics are susceptible to attack because some information is leaked from the biometric template [2]. Moreover, the security against the attack is underestimated [3]. Lopresti and Raim proposed a generative model to attack a handwriting authentication system [13]. The basic units of user's handwriting samples were manually segmented and then the corresponding units were concatenated to form a user' pass-phrase. Next, a feature space search was performed within a predetermined time limit of 60 seconds. As a result, their attack succeeded 49% of the time. In

later research, Ballard et al. expanded the work in [13]. In their work [3], they conducted a series of attacks using human-based and generative models to attack a handwriting authentication system. For the human-based attack, they used trained imposters who were allowed to select and replay real time renderings of a target user's pass-phrase in the experiments. The authors showed that the FAR of the trained imposters when compared to the untrained counterparts significantly increased. For the generative model, the results showed that the generative attack match or exceed the effectiveness of forgeries rendered by the trained imposters.

There are a number of successful attacks against speaker verification [14], [18], [15], [25], [11]. Masuko et. al. [14] presented an attack model using synthetic speech. An HMM-based speaker verification with an FAR (False Acceptance Rate) of 0% for human impostors was used as a baseline. They used an HMM-based speech synthesizer to create synthetic speech. The results showed that the FAR against the HMM-based synthesized speech increased to 70% by using only 1 sentence as training data. However, De Leon et. al. proposed a technique to detect the synthesized speech [5]. The fact that the HMM-based synthesizer will always produce the same optimal waveform in term of likelihood score was exploited to detect the synthetic speech. With this step, the speaker verification system was able to prevent some adversaries utilizing synthesized speech.

Yee, Wagner and Tran [25] reported the attack on a GMM-based speaker verification system against human impostors. Two people, male and female, played the roles of imitators against 138 speakers in the YOHO database. At first, the imitator was required to speak the same utterance in the YOHO database to calculate the similarity score between the imitator and the same-gender subject in the YOHO database. Among 138 speakers, the closest, intermediate, and furthest speakers have been selected by similarity scores. Finally, each imitator tried to mimic 3 target speakers in the database. The recordings were conducted on four sessions. The authors performed recordings on multiple sessions to consider the improvement of the imitator after each session. During each session, the imitators listened to each utterance once at a time and then repeated that utterance with their voices. The best result of the female imitator was accepted 30% of the time by the system, while the male imitator achieved a 35% acceptance rate.

3. Datasets

We will evaluate the recognition performance using an Equal Error Rate (EER) which is the rate at which a False Acceptance Rate (FAR) and a False Rejection Rate (FRR) are equal. The FAR is the percentage of the time that the system accepts the wrong speaker or one who is not authorized to access the system. In the same way, the FRR is the percentage of the time that the system rejects the authorized speaker. Two datasets are used in experiments: The MIT mobile device speaker verification corpus Dataset (MDS) [24] and The Lehigh quiet environment speaker verification Dataset (LDS). The MDS is a public dataset available by MIT. The LDS is our dataset collected over a month period.

3.1 The MIT mobile device speaker verification corpus

This dataset was collected from 48 speakers (22 females and 26 males). The utterances were recorded in three acoustic environments: office, lobby, and street intersection via two types of microphones: external earpiece headset and built-in mobile device. The dataset consists of two sets: a set of enrolled users and a set of dedicated imposters. For the enrolled set, speech data was collected over two sessions on separate days (20 minutes for each session). For the imposter set, users participated in a single 20 minutes session. There are six lists of pass-phrases that were varied by three environments and two types of microphones. We select the first list to our experiment because it provided pass-phrases that were said by the same speaker multiple times under the same environment (office). So, we can use this list in the training and the testing phase.

3.2 The Lehigh quiet environment speaker verification dataset

This dataset contains 4,320 recordings collected on a laptop computer via an external earpiece headset microphone from six male speakers during several rounds. The data collection was taken in the graduate study room that can be referred to as quiet environment. In the first round, the subjects were asked to say their five pass-phrases. Each pass-phrase was uttered 10 times. In addition, they were asked to say 270 short sentences to make a speech corpus. In the second round, they were asked to say their same set of pass-phrases. Each was uttered five times. Furthermore, they were asked to say other subjects' pass-phrases. Each was uttered five times. Lastly, they were asked to imitate the other subject pass-phrases by listening to the pass-phrases that we replayed to them. Each pass-phrase was uttered five times. By listening to imitated pass-phrases, we selected the best imitator for the third round. The best imitator was asked to mimic the target speaker's pass-phrases. Each pass-phrase was uttered five times.

4. Experimental Setup

4.1 Speaker verification models

We use a low-pass digital filter with a cut-off at 4 kHz to strip the higher frequencies from the signal. The next step is pre-emphasis, which is the process to raise the Signal to Noise Ratio. The signal is pre-emphasized by passing the signal to a first order digital filter $H(z) = 1-\alpha z^{-1}$, where we set $\alpha = 0.98$. Framing is the next step. The signal is

framed into the short time analysis interval. Each frame is multiplied by a window function (Hamming) to reduce abrupt changes at the start and the end of each frame. These frames have to be overlapped properly. The length of each frame is usually around 30 msecs; this length would yield good results for speech processing with 10 msecs overlap [10]. For the sampling rate of 8 kHz, we use 240 samples per frame that are shifted every 80 samples.

We use six utterances in the MDS and LDS for training. A decision threshold is estimated based on inter-speaker scores. For our setting, let mean and standard deviation of the inter-speaker score be μ and σ, we set the decision threshold $\theta = \mu - c\sigma$ where *c* is some constant that minimize the error rate.

For all constructions, we use 13 order Mel-Frequency Ceptral Coefficient (MFCC) [6] for training and verification.

4.1.1 Dynamic Time Warping (DTW)

Given two time sequences of feature vectors, we have to find a warping function which minimizes the distance between the two feature vectors. Let $A = a_1, a_2, \ldots, a_i, \ldots, a_M \quad and \quad B = b_1, b_2, \ldots, b_j, \ldots, b_N$ be the two sequences of feature vectors to be compared. The warping function can be represented by a sequence of lattice points on the plane, $L = l_1, l_2, \ldots, l_k, \ldots, l_K, \quad l_k = (i_k, j_k)$.

Let $d(l_k)$ be a cost function which is defined as the distance between a_{i_k} and b_{j_k}. The overall cost function, $D(L)$, can be determined by the following equation [17].

$$D(L) = \sum_{k=1}^{K} d(l_k) \tag{1}$$

We have to find the minimum-cost path from point (1,1) to point (M, N). The minimum-cost warping path can be efficiently determined by using Dynamic Programming (DP) which satisfy some necessary constraints such as in [20].

For DTW, we use the first utterance as the reference signal and perform DTW to the rest. The averaged result is stored as the matching template. The distance between an input and the matching template is determined by using the Euclidean distance. The system decides whether to accept or reject the speaker by comparing the Euclidean distance to the decision threshold.

4.1.2 Vector Quantization (VQ)

For VQ, the acoustic models of speakers are created by partitioning a collection of acoustic feature vectors to *C* clusters [22]. Each cluster is represented by a mean vector or centroid denoted by c_i for *i = 1, …, C*. In literature, a set of controid $\mathcal{C} = c_1, \ldots, c_C$ are referred to as a codebook. For verification, given an input vector *X = {x_1, …, x_m}* , the quantization distortion [12] for speaker *j* can be calculated by summing the nearest distance in the codebook ($\mathcal{C}_j$). More precisely, the distortion of the vector x_k from $\mathcal{C}_j$, $d(x_k, \mathcal{C}_j)$, is given by equation 2 where *$d(x_k, c_i)$* is a distance between x_k and c_i.

$$d(x_k, \mathcal{C}_j) = \arg\ min_{c_i \epsilon \mathcal{C}_j}\ d(x_k, c_i) \tag{2}$$

Hence, the distortion of *X* from $\mathcal{C}_j$ is determined by the following equation.

$$D(X, \mathcal{C}_j) = \frac{1}{m} \sum_{k=1}^{m} d(x_k, \mathcal{C}_j) \tag{3}$$

For our setting, the K-means clustering is used to quantize the training vectors. We investigate the performance of VQ in our datasets by setting the number of codebooks to 10, 20, 30, 40, and 50. The performance with 30 codebooks yields the best results. Therefore, we set *C = 30*. The distance between an input vector and the nearest centroid is determined by using the Euclidean distance. The system decides whether to accept or reject the speaker by comparing the distance to the decision threshold.

4.1.3 Gaussian Mixture Models (GMM)

The GMM model consists of a finite number of Gaussian distributions parameterized by their priori probability π_j, mean vectors μ_j, and covariance matrices Σ_j [19]. In this experiment, we use nodal covariance matrices. We initialize the speaker models using the K-means clustering, then the parameters are estimated by using the EM algorithm [8]. Given an input vector *X = {x_1, …, x_m}*, the matching score for GMM is given by the log-likelihood of the GMM L in the following equation where λ_j = (π_j, μ_j, Σ_j) and $\lambda_{j\prime}$ = ($\pi_{j\prime}$, $\mu_{j\prime}$, $\Sigma_{j\prime}$) are the model of speaker *j* and the background model of speaker *j*.

$$L = log\ p(X|\lambda_j) - log\ p(X|\lambda_{j\prime}) \tag{4}$$

The training utterances of all speakers except speaker *j* are used to create the background model and the rest is used to create the speaker model of speaker *j*. We use the GMM mixture order = 10 for the reason similar to the setting of the VQ. The system decides whether to accept or reject the speaker by comparing the log-likelihood to the decision threshold.

4.2 Attack models

We investigate two types of attack: human and algorithmic. We vary the adversary knowledge by making three different assumptions in the human case and two different assumptions in the algorithmic case, for a total of five classes of attacks.

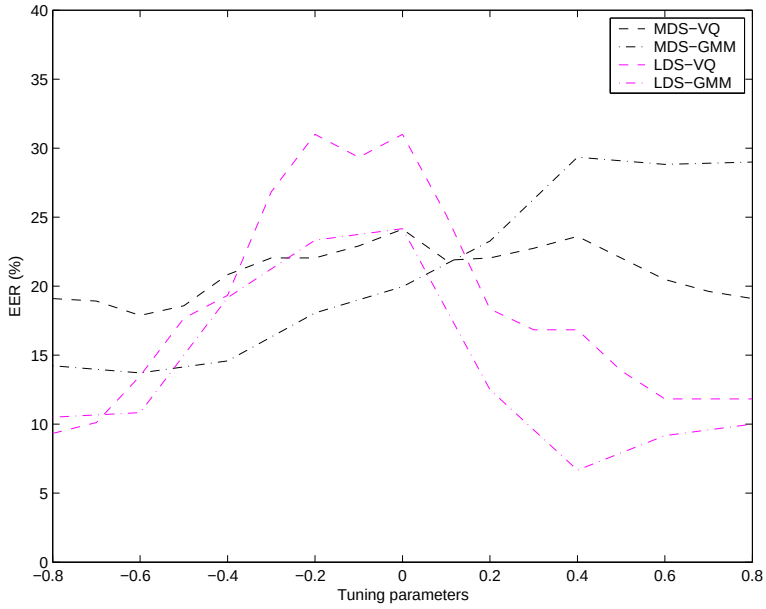

Fig. 1: The error rates (EERs) of re-generated pass-phrases for VQ and GMM system by varying κ.

4.2.1 The human type with assumption I (H-I)

We assume that the attackers do not know the authentic speakers and their pass-phrases. We evaluate the error rate of the authentic speakers compared with the adversary (*naive*) who say the random pass-phrase with different phonetic content than the actual pass-phrase.

4.2.2 The human type with assumption II (H-II)

We assume that the attackers know the pass-phrase and say the actual pass-phrase. In this experiment, the adversary (*imposter*) say the actual pass-phrase without listening to the target pass-phrase. All other subjects except the authentic speaker will be the adversaries.

4.2.3 The human type with assumption III (H-III)

We assume that the attackers know the pass-phrase and are acquainted with the authentic speaker. Then, they try to mimic the target pass-phrase. In our experiment, we re-play the pass-phrases of target speakers to the adversary (*informed imposter*) and then the informed imposter repeats the pass-phrases. Note that we use the term "informed" instead of "skilled" because the attackers just have been given useful information for creating a forgery. In addition, "skill" means that someone has demonstrated a proven talent; we have done nothing to prove that the test subjects actually have a real talent for forgery.

4.2.4 The algorithmic type with assumption I (A-I)

We assume that the attackers know the pass-phrase and have acquired some voice samples of the target user. Then, they synthesize the pass-phrase. In this work, we use HMM-based speech synthesizer to create *synthetic pass-phrases* [23]. We use 270 phrases in the first round in the LDS dataset for training the speaker-dependent models to synthesize pass-phrases. The phonemes in each phrase are labeled to form HMM models of each phoneme in the training phrase. The HMM models are parameterized by spectrum (MFCC) and excitation (F0, and duration) parameters. To synthesize the sound, the pass-phrase to be synthesized is analyzed then the phoneme HMM models are concatenated based on unit clustering. Finally, the concatenated HMM models output the parameters to synthesize the sound by passing these parameters to the synthesis filter. For each speaker, five pass-phrases are synthesized corresponding to their pass-phrases. For all processes in synthesizing the sound, we set the sampling rate at 8000 Hz.

4.2.5 The algorithmic type with assumption II (A-II)

We assume that the attackers know the pass-phrase and have acquired the template of the target user. Moreover, they know the system's construction and use this information to create *re-generated pass-phrases*.

We refer the algorithmic type attacker (A-I and A-II) to as an *informed adversary*.

Attack against DTW template: We store 13 order MFCCs of the training utterances as the matching template. Hence, we have to transform this template to a signal. In this experiment, we use Auditory Toolbox [21] to derive MFCCs. Then, we use the same tool to reconstruct the speech signal used as a forgery.

Attack against VQ and GMM template: For VQ and

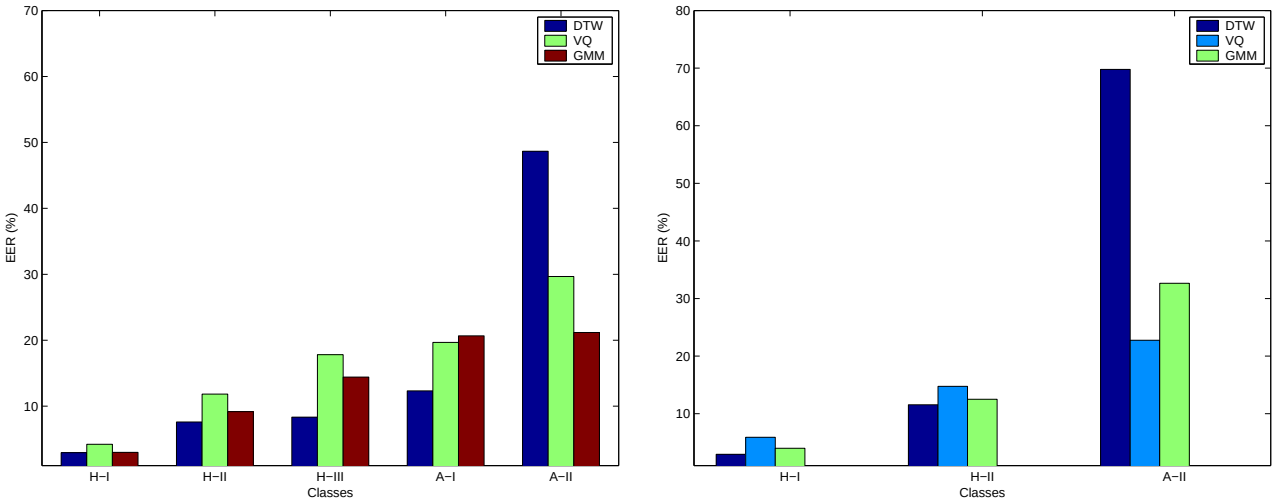

Fig. 2: The EERs against various attacks and models in the LDS (left) and MDS (right)

GMM, the verification consists of two units. The first is speech recognition which is used to check whether user utter the registered pass-phrases. The second is speaker verification which aims to determine whether the person is who he/she claims to be. For the speech recognition unit, the implementation is based on DTW. A set of pass-phrases from the speakers is used to create templates of *W* classes. Each class consists of *R* reference templates. An input vector will be aligned to the same range of the set of reference templates in each class. Hence, we employ k-nearest neighbor for classification [7]. Next, the unknown spoken input will be classified into one of *W* classes. For our datasets, *W* is 10 for the MDS and 30 for the LDS. The accuracy of the recognition unit is 90.58% and 94.64%.

The VQ template consists of a codebook $\mathcal{C}$ and a decision threshold θ for each speaker. These parameters will be used to calculate the distortion of a set of input vectors *X* (a set of range m). The system will accept the speaker, if the distortion of the *X* is lower than the threshold. Hence, we will search a set of vector $x_i \in X$ that yield a distortion close to the decision threshold.

Let the distortion of x_i from $\mathcal{C}$ be $d(x_i, \mathcal{C})$. We want to select a set of vectors $s = \{x_i | d(x_i, \mathcal{C}) < T_a, i = 1, 2, \ldots, n\}$ where $n \leq m$ and T_a is the appropriate threshold to re-synthesize the pass-phrase. Basically, the verification performance will be degraded if T_a is high. On the other hand, by setting T_a too low, it will degrade the speech recognition. Hence, the appropriate threshold will be determined by experimentation. More precisely, we will set $T_a = \theta + \kappa\theta$ where $\kappa \in [-a, a]$ is *a tuning parameter*. Then, we select κ which yields the best result. A possible problem is the case of a null set of s. In this case, we use the source's vectors (*X*).

For the GMM attack, the priori probability π_j, mean vectors μ_j, and covariance matrices Σ_j are used to calculate the log-likelihood of the input vectors. We will select a set of vectors based on the decision threshold of log-likelihood.

5. Experimental Results

For the LDS, we use five pass-phrases from each speaker in our experiment, a total of 5*6 = 30 different pass-phrases. Six recordings from the first round are used to train the system. Five recordings from the second round are used for verification. We randomly select 25 other pass-phrases from other speakers that do not correspond to the verification pass-phrase to evaluate H-I's trial. Five recordings of the same pass-phrase uttered by other speakers in the second round are used to evaluate H-II's trial, in total of 5*5=25 recordings for each pass-phrase. For H-III's trial, five mimicked recordings are used. The synthesized pass-phrase is used for A-I's trial. For A-II's trial, we use five recordings from H-II's trial as the sources of acoustic features and change them to target pass-phrases.

For the MDS, we use six recordings to train the systems. Two recordings are used for verification. The number of H-II's pass-phrases that are available in the dataset varies from one to six. Hence, these pass-phrases are used as sources of acoustic features for A-II. For H-I's trial, we use six pass-phrases from other speakers that are different from the verification pass-phrase. For the other classes, we do not have enough voice samples to synthesize reasonably high quality sound and we do not have mimicked utterances. Hence, we do not investigate H-III and A-I.

The results in Figure 1 illustrate the error rates of re-generated pass-phrases by varying $\kappa \in [-0.8, 0.8]$. Maximum points of each plot optimize the trade off between the recognition and verification performance of the systems. Thus, we set κ to the maximum point for each system.

Figure 2 depicts the graphical results of EERs against the various attacks for the LDS and MDS. It is clear that the

Table 1: FARs (%) of speaker verification systems against various attacks using decision thresholds at operating points of imposters (H-II).

Datasets	Attack models	DTW	VQ	GMM
LDS	H-I	0.27	3.53	2.22
	H-II	7.20	11.56	8.89
	H-III	8.67	25.47	24.05
	A-I	20.00	26.67	60.00
	A-II	**90.00**	**55.00**	**65.00**
MDS	H-I	0.00	4.08	2.08
	H-II	11.86	16.40	13.12
	A-II	**100.00**	**47.22**	**89.93**

informed adversary utilizing the template information (A-II) is the most successful adversary in gaining access to all systems. In particular, in the DTW system.

The results of A-II from the MDS and LDS seem to conflict (Figure 2). The A-II algorithm for the GMM did better in the MDS, but in the LDS the result is inverse. This may be possible for two reasons. First, we vary the tuning parameters roughly. Thus, the better value of κ may be missed. The other reason is that we just utilize imposters' pass-phrases (H-II) in the case of a null set of s. Hence, these pass-phrases may affect the results. For the other attack models, the EERs of the DTW are the lowest. In particular, for the H-III and A-I in the LDS, the EERs of the DTW are noticeably lower than the EERs of the VQ and GMM. These results suggest that the DTW will yield a good performance if the template is protected properly. Thus, the template protection is the critical issue for the DTW approach.

Assuming that the practitioners do not take the informed imposter and adversary (H-III, A-I, and A-II) into account, a decision threshold may be determined to be at an operating point of H-II. We further assume that the systems do not check whether the pass-phrase is correct because for text-dependent speaker verification systems if the pass-phrase is incorrect, the matching score will be greater than the threshold and eventually be rejected. The results are summarized in Table 1 which illustrates the error rates (FAR) of various attacks. The figures of H-I and H-II in the table reflect the standard (traditional) evaluation of biometric authentication systems. Beyond the standard evaluation, the FARs of other attack models are very high. In particular, the FARs of A-II are the highest.

6. Conclusions

In this work, we have shown that the adversary can exploit the DTW, VQ and GMM template and use them to attack the systems. We developed an algorithm to re-generate the pass-phrases that can be used in remote or on-line authentication. We compared our algorithmic attack with the traditional (human imposters) and the more sophisticated attack (an adversary utilizing a synthetic pass-phrase). The EERs of the re-generated pass-phrases were better than the other attack models. Then, we have demonstrated that the traditional approach to evaluate the security of speech biometric speaker verifications was insufficient. The results indicated that the FARs of other attack models beyond the traditional approach were very high.

We hope that these results will provide an important warning for researchers when attempting to demonstrate the security of speech biometric systems. For future work, we are considering ways to address the weaknesses we have identified in this work.

7. Acknowledgments

The authors would like to thank Jim Glass who provided us with the MIT mobile device speaker verification corpus for this research. We would also like to thank Candice Quinones, an adjunct of English as a Second Language (ESL) at Lehigh University, for her review of the previous version of this manuscript.

References

[1] L. Ballard. *Robust technique to evaluate the security of biometric*. PhD thesis, The Johns Hopkins University, Baltimore, Maryland, March 2008.

[2] L. Ballard, S. Kamara, and M. K. Reiter. The practical subtleties of biometric key generation. In *Proceedings of The 17th Annual USENIX Security Symposium*, pages 61-74, San Jose, CA, August 2008.

[3] L. Ballard, D. Lopresti, and F. Monrose. Forgery quality and its implications for behavioral biometric security. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics (Special Edition). 37(5):1107-1118, October 2007.

[4] J. P. Campbell. Speaker recognition: a tutorial. In *Proceedings of The IEEE* Vol.85 No.9, pages 1437-1426, September 1997.

[5] P. L. De Leon, V. R. Apsingekar, and J. Yamagishi. Revisiting the security of speaker verification systems against imposture using synthetic speech. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1798-1801, Dallas TX, USA, 2010.

[6] J.R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. Macmilland Pub. Co., New York, 1993.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. New York: Wiley, 2001.

[8] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):1-38, 1997.

[9] D. Feldmeier and P. Karn. UNIX password security-ten years later. In *Advances in Cryptology-CRYPTO'89*, pages 44-63, Springer Verlag, London, UK, 1989

[10] S. Furui. *Digital speech processing, synthesis and recognition*. Marcel Dekker, Inc., New York, 2001.

[11] Q. Jin, A. Toth, A. Black, and T. Schultz. Is voice transformation a threat to speaker identification? In proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2008), pages 4845-4848, April 2008.

[12] T. Kinnunen.*Spectral Featrues for Automatic Text-Independent Speaker Recognition*. Licentiate thesis, Department of Computer Science, University of Joensuu, Findland December 2003.

[13] D. Lopresti, J. Raim. The effectiveness of generative attacks on an online handwriting biometric, In *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, pages 1090-1099, NY, USA, July 2005.

[14] T. Masuko, T. Hitotsumatsu, K. Tokuda and T. Kobayashi. On the security of HMM-based speaker verification systems against imposture using synthetic speech. In *Proceedings of the European Conference on Speech Communication and Technology*, Vol.3, pages 1223-1226, Budapest, Hungary, September 1999.

[15] T. Masuko, K. Tokuda, and T. Kobayashi. Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Proceedings of the International Conference on Spoken Language Processing*, Vol. 3, pages 302-305, Beijing, China, October 2000.

[16] K. Nandakumar, A. Nagar, and A. K. Jain. Hardening fingerprint-based fuzzy vault using password, In *proceedings of 2nd International Conference on Biometrics (ICB)*, pages 927-937, Seoul, South Korea, August 2007.

[17] T. W. Parsons.*Voice and Speech Processing*. McGraw-Hill, New York, 1987.

[18] B. L. Pellom and J. H. L. Hansen. An experimental study of speaker verification sensitivity to computer voice altered imposters. In *Proceedings of the 1999 International Conference on Acoustics, Speech, and Signal Processing*, March 1999.

[19] D. A. Reynolds, T. F. Quatieri, and Robert B. Dunn. Speaker verification using adapted guassian mixture models. *Digital signal processing*. 10(1-3), pages 19-41, 2000.

[20] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transaction on acoustics, speech, and signal processing*, ASSP-26 (1): 43-49, February 1978.

[21] M. Slaney. Auditory Toolbox. *Interval Technical Report# 1998-010*, 1998.

[22] F. K. Soong, A. E. Rosenberg, B. Juang, and L. Rabiner. A vector quantization approach to speaker recognition. AT&T Technical Journal 65. pages 14-26. 1987.

[23] K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to English. In *Proceedings of IEEE Speech Synthesis Workshop*, 2002.

[24] R. H. Woo, A. Park, and T. J. Hazen. The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments. In *Proceedings of Odssey, The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006.

[25] W. L. Yee, M. Wagner and D. Tran, Vulnerability of speaker verification to voice mimicking. In *Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 145-148, Hong Kong, October 2004.