SAND2014-1493C

PERFORMANCE OF I-VECTOR SPEAKER VERIFICATION AND THE DETECTION OF SYNTHETIC SPEECH

Richard D. McClanahan

Sandia National Laboratories
Albuquerque, N.M., U.S.A.
rmcclan@sandia.gov

Bryan Stewart, Phillip L. De Leon

New Mexico State University
Klipsch School of Elect. and Comp. Eng.
Las Cruces, N.M., U.S.A.
{brystewa,pdeleon}@nmsu.edu

ABSTRACT

In this paper, we present new research results on the vulnerability of speaker verification (SV) systems to synthetic speech. Using a state-of-the-art i-vector SV system and evaluating with the Wall-Street Journal (WSJ) corpus, our SV system has a 0.00% false rejection rate (FRR) and 1.74×10^{-5} false acceptance rate (FAR). When the i-vector system is tested with state-of-the-art speaker-adaptive, hidden Markov model (HMM)-based synthetic speech generated from speaker models derived from the WSJ journal corpus, 22.9% of the matched claims are accepted highlighting the vulnerability of SV systems to synthetic speech. We propose a new synthetic speech detector (SSD) which uses previously-proposed features derived from image analysis of pitch patterns but extracted on phoneme-level segments and which leverages the available enrollment speech from the SV system. When the SSD is applied to human and synthetic speech accepted by the SV system, the overall system has a FRR of 7.35% and a FAR of 2.34×10^{-4} which is lower than previously-reported systems and thus significantly reduces the vulnerability.

Index Terms— Speech synthesis, Speaker recognition, Security

1. INTRODUCTION

Recently, text-to-speech (TTS) systems or speech synthesizers have advanced to the point where they can be trained to a particular person's voice or *target*. Such training, using state-of-the-art speaker-adaptive, hidden Markov model (HMM)-based speech synthesizers, now only requires relatively small amounts of non-ideal speech which can be acquired in a variety of ways including through online media [1], [2]. The speech is then used to adapt an average (derived from other speakers) or a background (derived from one speaker) synthesizer model yielding a target model. With the target model, arbitrary speech utterances can be synthesized in real-time in an acoustically similar fashion to the target voice.

In 1999, Masuko et. al., showed in a limited study that a speaker verification (SV) system would accept identity claims based on synthetic speech [3]. This work showed that synthetic speech provides a potential means for an adversary to potentially gain system access when speech is used to authenticate the identity claim.

In the decade after Masuko's work, both SV and TTS systems improved dramatically. Beginning in 2010, De Leon, et. al. reinvestigated the vulnerability of SV systems to synthetic speech using state-of-the-art systems [4–6]. In the most recent work, [6], we examined SV systems based on the Gaussian mixture model-universal background model (GMM-UBM) [7] and support vector machine (SVM) using GMM supervectors [8]. We used 283 speakers from the Wall Street Journal (WSJ) corpus which was partitioned into non-overlapping datasets for SV enrollment ($\approx$ 90 s signals), SV testing ($\approx$ 30 s signals), and TTS training (varying amounts from 73 s to 27 minutes) [6]. Although the WSJ corpus is not a standard corpus for SV research, it is one of the few corpora that provides several hundred speakers and sufficiently long signals required for constructing each of the components within the TTS, SV, and SSD systems [9]. We then created speaker-adaptive, HMM-based synthetic test speech for each of the WSJ speakers. As we demonstrated, under human speech the EERs were 0.284%, 0.002% for the GMM-UBM, SVM system respectively. However, when subjected to synthetic speech, the matched claim rate i.e. a synthetic signal matched to a targeted speaker and an identity claim of that same speaker, was over 81% for each of the SV systems demonstrating the vulnerability of SV systems to synthetic speech.

In addition, vulnerabilities of SV systems to voice conversion, i.e. conversion of a speaker's voice into a target voice, have also been reported as far back as 1999 [10]. Using state-of-the-art joint density GMM (JD-GMM) and unit-selection techniques, researchers in [11, 12] have recently evaluated the vulnerability of several SV systems to voice conversion. In particular, a joint factor analysis (JFA) SV system (precursor to the i-vector system) was shown to have FARs increase from 3% to over 15% under JD-GMM voice-converted speech [11]. In [12], researchers investigated both text-independent and text-dependent GMM-UBM SV systems. For the text-independent system, the EER (average of male and female tests) increases significantly under voice-converted speech from 16.2% to 28.8%, 26.7% using unit-selection, JD-GMM techniques, respectively [12]. For the text-dependent system, the EER (average of male and female tests) drops from 5.7% to 3.6%, 3.2% using unit-selection, JD-GMM techniques, respectively [12]. These results collectively illustrate that voice-converted speech also poses a threat to SV systems.

Because of the vulnerability of SV systems to both synthetic and voice-converted speech, researchers have proposed various counter-measures and detectors [13]. In 2001, Satoh et. al. proposed a method to detect synthetic speech based on the average inter-frame difference of log-likelihood (IFDLL) [14]. However, in 2010 with state-of-the-art synthetic speech, it was demonstrated that IDFLL could no longer discriminate between human and synthetic speech [15]. In [5], a discriminator based on relative phase shift (RPS) was proposed for detecting synthetic speech and it was shown that the acceptance rate of synthetic speech, matched claimants (matched claim rate) could be lowered from over 81% to 2.5% with less than a 3% drop in the acceptance rate for human speech, true claimants. However, the detector was found to be sensitive to the

vocoder used: the same vocoder used by the impostor must be used to train the system which is not a general solution.

In [16], the authors proposed using additional features extracted from the phase spectrum in order to detect voice-converted speech. One feature, cos-phase, unwraps the phase spectrum and applies a cosine function to normalized and a discrete cosine transform (DCT) to reduce dimensionality of the feature vector [16]. A second feature, modified group delay function (MGDF), is based on the group delay of a smoothed power spectrum and parameters used to emphasize the spectral fine structure [16]. Evaluation with a JFA SV system using the NIST 2006 SRE corpus and GMM-converted speech shows a baseline EER of 16.8% which is reduced to 6.60% with the cos-phase feature and 9.13% with the MGDF feature. Evaluation using unit-selection converted speech shows a baseline EER of 15.4% which is reduced to 3.9% with the cos-phase feature and 4.6% with the MGDF feature.

In [17], we proposed an utterance-level classifier for synthetic speech detection based on features extracted from image analysis of pitch patterns that did not require synthetic models matched to humans in the SV system or any a priori information regarding speech synthesizers. These features which include the mean pitch stability, mean pitch stability range, and jitter were found to provide good discrimination between human and synthetic speech. Follow on work in [18] proposed a word-level classifier using the same feature set as in [17]. However, the results presented in [17] used an utterance-level likelihood classifier and a different set of training and testing corpora than what was presented in [18]. The utterance-level classifier used in [17] was re-evaluated with the training and testing corpora presented in [18] and the results showed 96%, 92% classification accuracy for human, synthetic speech, respectively. The results in [18] for the proposed word-level maximum likelihood classifier using the Bhattacharyya weighted mean feature vector, showed improved classification accuracy of 98%, 98% for human, synthetic speech, respectively.

This paper reports our research on the performance of SV systems under synthetic speech and a proposed new method to detect synthetic speech. First, we have implemented a state-of-the-art i-vector SV system [19, 20] and evaluated it using synthetic speech. As we will show, synthetic speech continues to pose a threat to SV systems. Second, we propose a new synthetic speech detector based on features extracted from image analysis of binary pitch patterns from phoneme segments. Unlike prior work, this new detector does not require a parallel human/synthetic corpus for training but rather leverages the available enrollment speech (assumed to be human) used in SV training and thus is more general and practical.

This paper is organized as follows. In Section 2, we briefly describe our implementation of an i-vector SV system, system development, training and evaluation. In Section 3, we briefly describe the proposed synthetic speech detector based on features extracted from pitch patterns from phoneme segments. In Section 4, we provide the baseline evaluation of the SV system using the WSJ journal corpus and synthetic speech derived from this corpus as well as performance of the overall system with the synthetic speech detector. Finally, we conclude the article in Section 5.

2. SPEAKER VERIFICATION SYSTEM

For this paper, we use the state-of-the-art i-vector SV system described by Dehak, et. al. in [19, 20] and Garcia-Romero in [21]. We briefly describe the system, our implementation, system development, training, and evaluation.

2.1. System Development

Speech signals from the NIST 2004, 2005, 2006, and 2008 SREs were used in system development. This data was used for UBM training, total variability (TV) training, and probabilistic linear discriminant analysis (PLDA) parameter training. Our system is based on 40-dimensional feature vectors extracted as follows. We extract 20 mel-frequency cepstral coefficients (MFCCs), including the zeroth coefficient, using a 25 ms Hamming window with 10 ms advance. MFCCs are then RASTA filtered and we compute the log energy. The Δ of the features are computed and appended to each feature. Short-time mean and variance normalization is applied to the feature vector using a 3 s window and vectors corresponding to silence are finally removed.

The steps involved in system development include training the GMM-UBM, estimation of the TV matrix, and parameter estimation for PLDA. We construct a single gender-independent GMM-UBM with 1024 components and diagonal covariance matrices using the Expectation Maximization (EM) algorithm. The GMM-UBM is trained using feature vectors from 30,000 randomly-selected utterances from the NIST SRE corpora. We estimate a single gender-independent TV matrix with a rank of 400 using the EM algorithm. The TV matrix is trained from feature vectors extracted from the NIST SRE corpora. PLDA parameters were then estimated using a varying number of utterances per speaker. In our SR system, we did not attempt to prevent overlap in training utterances between different system components nor did we attempt to limit our data to a particular channel type such as telephone or microphone.

2.2. System Training

After system development, i-vectors for target speakers are extracted from the WSJ corpus. In our prior work using the WSJ corpus, we chose the pre-defined official training data set, SI-284, that includes 283 speakers from both WSJ0 and WSJ1 as material data [6]. For this research, we use all 340 WSJ speakers and training signals which are approximately 180 s in length. After the i-vector is extracted, we compensate by applying length normalization.

2.3. System Testing

Once the system has been trained, we extract an i-vector for the test speaker and use PLDA for scoring against the target i-vector. No score normalization–such as z-norm or t-norm–is applied to the PLDA output. We use WSJ test signals which were approximately 30 s in length.

2.4. Evaluation

The evaluation for human speech was designed so that each test utterance has an associated true claim and 339 false claims (impostors) yielding a total of 340^2 tests. Our system has a 0.00% false rejection rate (FRR) and 1.74×10^{-5} false acceptance rate (FAR), i.e. two false acceptances out of the 340×339 possible imposter tests. The low FRR and FAR are due to the ideal nature of the recordings in the WSJ corpus and the accuracy of the i-vector SV system.

3. SYNTHETIC SPEECH DETECTOR

Figure 1 illustrates a SV system with a countermeasure for synthetic speech. The test signal is applied to a synthetic speech detector (SSD) after the SV system has initially accepted the identity claim based on the signal [6]. If the SSD classifies the test signal as human

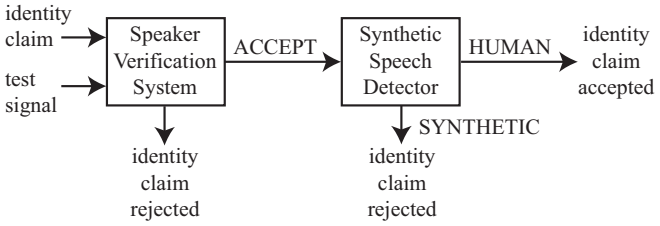

Fig. 1. Integrated speaker verification (SV) and synthetic speech detector (SSD) system. Only those claimants which are accepted by the SV system are passed to the SSD. If the SSD classifies the speech as human, the overall system accepts the claim.

speech, the claim is finally accepted, otherwise it is rejected. In this section, we briefly summarize the SSD used in this research which uses image-based pitch pattern features recently proposed in [17].

3.1. Image-Based Pitch Pattern Features

The pitch pattern, $\phi(t,\tau)$, is calculated by dividing the short-range autocorrelation function, $r(t,\tau)$ by a normalization function, $p(t,\tau)$ which is proportional to the frame energy [22]

$$\phi(t,\tau) \quad = \quad \frac{r(t,\tau)}{p(t,\tau)}. \tag{1}$$

Once the pitch pattern is computed, we segment into a binary pitch pattern image through the rule

$$\phi_{\text{seg}}(t,\tau) \quad = \quad \begin{cases} 1, & \phi(t,\tau) \geq \theta_t \\ 0, & \phi(t,\tau) < \theta_t \end{cases} \tag{2}$$

where θ_t is a threshold set to half the pitch pattern peak value at time t. We compute $\phi(t,\tau)$ for $2 \leq \tau \leq 20$ms and set $\theta_t = 1/\sqrt{2}$ for all t. An example pitch pattern image is shown in Fig. 2.

Extracting features from the pitch pattern is a multi-step process and includes 1) silence removal, 2) computation of the pitch pattern, and 3) image analysis. In the third step, image processing of the segmented binary pitch pattern is performed in order to extract the connected components, i.e. black regions in Fig. 2. The resulting connected components are then analyzed and used to compute mean pitch stability, μ_S and mean pitch stability range, R_c which are defined in [18] and are the elements of the feature vector,

$$\mathbf{x} \quad = \quad [\mu_S, R_c]\,. \tag{3}$$

3.2. Synthetic Speech Detector based on Phoneme-Level Features

Extending our prior work from [17, 18], we propose a new classifier for the detection of synthetic speech based on pitch pattern feature vectors extracted from *phoneme-level* segments and which also leverages enrollment speech used in training the SV system. The classifier is illustrated in Fig. 3. In the training stage, for each speaker enrolled in the SV system, we use an automatic speech recognizer (ASR) to parse the enrollment speech signal into phoneme-level segments, extract the pitch pattern feature vectors for each segment corresponding to a voiced phoneme, and store vectors for each unique phoneme.

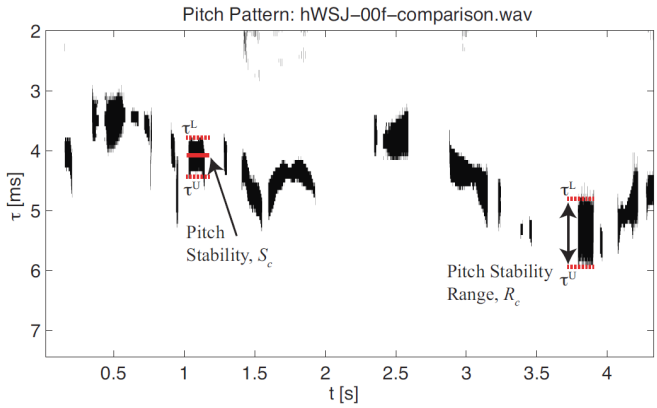

Fig. 2. Segmented binary pitch pattern image from a human speech signal [17]. The phrase is "The female produces a litter of two to four young in November." Pitch stability S_c, pitch stability range R_c, upper edge τ^{U}, and lower edge τ^{L} are denoted.

As shown in Fig. 3, the SSD parses the test signal into its phoneme segments using the ASR, pitch pattern feature vectors of the voiced phoneme segments are extracted, and the Mahalanobis distances are computed between the claimant speaker's and the corresponding target speaker's phoneme feature vectors. The mean of the Mahalanobis distances across the test utterance is computed and the claimant is classified as human if the distance is greater than a pre-defined threshold set for equal error rates; otherwise, if the distance is less than the threshold, the claimant speech is classified as synthetic. The WSJ speech used in SV enrollment was used for training the SSD and the WSJ human and synthetic speech used in testing the i-vector SV system (see next Section), was used to measure classifier (SSD) performance. The EER (of the stand-alone SSD without the SV system) is found to be 32%.

Although the pitch pattern feature vectors are extracted from segments corresponding to voiced segments, our research has shown that the unvoiced phonemes such as the unvoiced plosives, unvoiced fricatives, and affricates can be used to further improve the SSD accuracy. If the claimant speaker was identified as human but a majority of the aforementioned unvoiced phonemes in the test utterance do not yield any connected components in the pitch pattern image, then the claimant is classified as synthetic. With the use of the unvoiced phoneme segments, the EER (of the stand-alone SSD without the SV system) is further reduced to 12%. Figure 4 shows the detection error tradeoff (DET) curve for the SSD which uses both voiced and unvoiced phonemes for the classifier.

4. EXPERIMENTS AND RESULTS

4.1. Corpora

For this research, we use all 340 WSJ speakers and training signals which are approximately 180 s and test signals which are approximately 30 s. The WSJ corpus was used to construct 340 different speaker models using a speaker-adaptive, HMM-based speech synthesis system, H Triple S (HTS) [23]. These WSJ HTS speaker models were used in Festival to generate the synthetic WSJ speech.

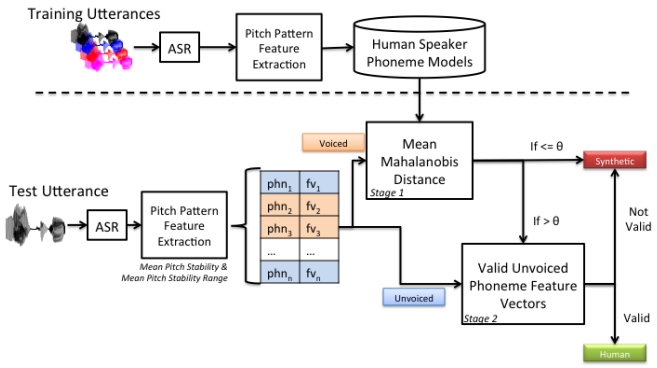

Fig. 3. Overview of the synthetic speech detector (SSD). In the training stage, enrollment speech from the SV system is used to provide pitch pattern feature vectors for each speaker's unique phonemes. The SSD classifies speech as human or synthetic by computing distances from feature vectors extracted from the claimant's speech and the target speaker's feature vectors.

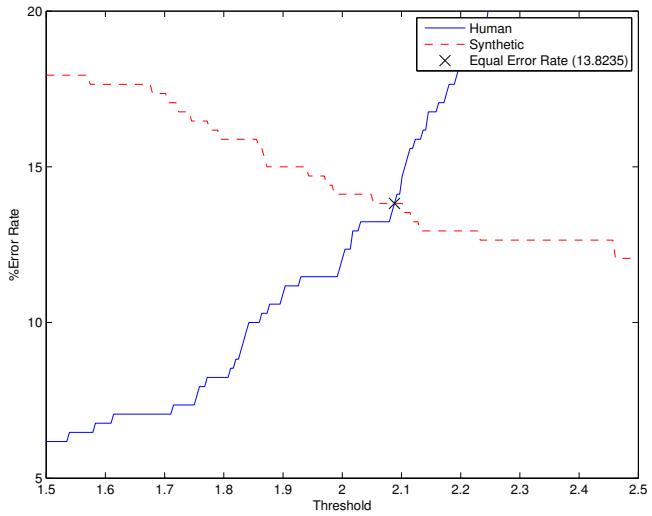

Fig. 4. Detection error tradeoff curve for the synthetic speech detector which uses features extracted from both voiced and unvoiced phoneme segments.

4.2. Matched Claim Acceptance Rate

The i-vector SV system was evaluated using synthetic speech derived from models based on the WSJ corpus. In this work, we find a matched claim rate of 22.9% (78/340). Although this a significant improvement over prior results with the GMM-UBM and SVM systems where we found over 81% of synthetic speech was accepted, it nevertheless demonstrates the continued vulnerability of state-of-the-art SV systems to synthetic speech.

4.3. Performance of Integrated SV and SSD System

The integrated SV and SSD system is illustrated in Fig. 1. For the WSJ corpus, we have from Section 2.4 that the SV system accepts 340 true claimants and 2 impostors. Using 340 synthetic speakers (matched to WSJ true claimants), we have from Section 4.2 that the SV system accepts 78 synthetic speech signals and rejects 262 synthetic speech signals. Of the 340 true human claimants which the SV system accepts, the SSD correctly classifies 315 of these as human and 25 incorrectly as synthetic; both impostors are classified as human. Of the 78 synthetic matched claimants which the SV system accepts, the SSD incorrectly classifies 25 of these as human and 53 correctly as synthetic. Therefore, the false rejection rate (FRR) of the overall system is $25/340 = 7.35\%$ and the false acceptance rate (FAR) is $(2 + 25)/(340 \cdot 339 + 340) = 2.34 \times 10^{-4}$. These results are an improvement over the system presented in [6] since the classifier does not require synthetic speech matched to each speaker enrolled in the SV system. These results are more accurate than the systems presented in [17] and [18] due to leveraging the speech used in enrollment for the SV system in order train the SSD and thus more accurately classify the test speech as human or synthetic.

5. CONCLUSIONS

In this paper, we have presented new results from our research into the vulnerability of speaker verification (SV) systems to synthetic speech. A state-of-the-art i-vector SV system was evaluated using 340 speakers from the WSJ corpus and shown to have a 0.00% false rejection rate (FRR) and 1.74×10^{-5} false acceptance rate (FAR). We evaluated the SV system using synthetic speech derived from models based on the WSJ corpus and found that 22.9% of the synthetic speech signals (matched to true human claimants) were accepted. This a significant improvement over prior results with the GMM-UBM and SVM systems where we found over 81% of synthetic speech was accepted.

We also have proposed a new synthetic speech detector (SSD) which uses previously-proposed features derived from image analysis of pitch patterns but extracted on phoneme-level segments and which leverages the available enrollment speech from the SV system. When the proposed SSD is integrated with the i-vector SV system, the overall system has a false rejection rate (FRR) of 7.35% and a false acceptance rate (FAR) of 2.34×10^{-4} which is lower than previously-reported systems and further reduces the vulnerability of to synthetic speech.

6. REFERENCES

[1] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[3] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. European Conf. Speech Communication and Technology (Eurospeech)*, 1999.

[4] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. IEEE*

Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2010, pp. 1798–1801.

[5] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 4844–4847.

[6] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of synthetic speech," *IEEE Trans. Speech, Audio & Language Process.*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Dig. Sig. Process.*, vol. 10, pp. 19–41, 2000.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[9] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. DARPA Speech and Language Workshop*, 1992, pp. 357–362.

[10] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1999, pp. 837–840.

[11] T. Kinnunen, Z. Z. Wu, K. A. Lee, F. Sedlak, E. .S Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 4401–4404.

[12] Z. Wu, A. Larcher, K. A. Lee, E. S. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints," in *Proc. Int. Speech Commun. Association (Interspeech)*, 2013.

[13] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and counter measures for automatic speaker verification," in *Proc. Int. Speech Commun. Association (Interspeech)*, 2013.

[14] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. European Conf. Speech Communication and Technology (Eurospeech)*, 2001, pp. 759–762.

[15] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, 2010, pp. 151–158.

[16] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Int. Speech Commun. Association (Interspeech)*, 2012.

[17] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Int. Speech Commun. Association (Interspeech)*, 2012.

[18] P. L. De Leon and B. Stewart, "Synthetic speech detection based on selected word discriminators," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013.

[19] 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. Int. Speech Commun. Association (Interspeech)*, 2009, pp. 1559–1562.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[21] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Proc. Int. Speech Commun. Association (Interspeech)*, 2011, pp. 249–252.

[22] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. Fundamentals*, vol. E88, no. 1, pp. 280–286, Jan. 2005.

[23] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio, and Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.