Evaluation of Speaker Verification Security and Detection of Synthetic Speech

Phillip L. De Leon, *Member, IEEE,* Michael Pucher, *Member, IEEE,* Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga

***Abstract*—In this paper, we evaluate the vulnerability of speaker verification (SV) systems to synthetic speech. The SV systems are based on either the Gaussian Mixture Model-Universal Background Model (GMM-UBM) or Support Vector Machine (SVM) using Gaussian supervectors. We use a Hidden Markov Model (HMM)-based text-to-speech (TTS) synthesizer, which can synthesize speech for a target speaker using small amounts of training data through model adaptation of an average voice or background model. Although the SV systems have a very low equal error rate (EER), when tested with synthetic speech generated from speaker models derived from the Wall-Street Journal (WSJ) speech corpus, over 91% of the matched claims are accepted. This result suggests a vulnerability in SV systems and thus a need to accurately detect synthetic speech. We propose a new feature based on relative phase shift (RPS), demonstrate reliable detection of synthetic speech, and show how this classifier can be used to improve security of SV systems.**

***Index Terms*—speaker recognition, speech synthesis, security**

I. Introduction

THE objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample [1]. Many investigations on the imposture problem as related to SV have been reported over the years as well as methods to prevent such impostures. The simplest imposture is playback of a voice recording for a targetted speaker and the well-known solution is a text-prompted approach [2]. In addition, the vulnerability of SV to voice mimicking by humans has also been examined in [3], [4]. On the other hand, advanced speech technologies present new problems for SV systems including imposture using speech manipulation of a recorded voice via analysis-by-resynthesis methods [5]–[7], voice conversion of the recorded voice [8]–[11], and diphone speech synthesis methods [7].

The use of synthesized speech potentially poses two related problems for SV systems. The first problem is confirmation of an acquired speech signal as having originated from a particular individual. In this case, the speech signal might be incorrectly confirmed as having originated from an individual when in fact the speech signal is synthetic. The second problem is in remote or on-line authentication where voice is used. In this case, a synthesized speech signal could be used to wrongly gain access to person's account and text-prompting would not present a problem for a text-to-speech (TTS) system. In both of these problems, the speech model for the synthesizer must be targeted to a specific person's voice. SV is also being used in forensic applications [12] and therefore security against imposture is also of obvious importance.

The problem of imposture against SV systems using synthetic speech was first published over 10 years ago by Masuko, et al. [13]. In their original work, the authors used an Hidden Markov Model (HMM)-based text-prompted SV system [2] and an HMM-based TTS synthesizer. In the SV system, feature vectors were scored against speaker and background models composed of concatenated phoneme models (not GMM-based models). The acoustic models used in the speech synthesizer were adapted to each of the human speakers [14], [15]. When tested with 20 human speakers, the system had a 0% False Acceptance Rate (FAR) and 7.2% False Rejection Rate (FRR); when tested with synthetic speech, the system accepted over 70% of matched claims, i.e. a synthetic signal matched to a targeted speaker and an identity claim of that same speaker.

In subsequent work by Masuko, et al. [16], the authors extended the research in two ways. First, they improved their synthesizer by generating speech using pitch information. Second, they improved their SV system by utilizing both pitch and spectral information. The pitch modeling techniques used in synthesis were the same used in the SV system. By improving the SV system, the authors were able to lower the matched claim rate for synthetic speech to 32%, however, the FAR for the human speech increased to 1.8%.

In the last 10 years, both SV and TTS systems have improved dramatically. Around the same time as Masuko's work, Gaussian Mixture Model-Universal Background Model (GMM-UBM) SV systems were first proposed [1]. Since this time, GMM-UBM based SV systems have produced excellent performance and have achieved EERs of 0.1% on the TIMIT corpus (ideal recordings) and 12% on NIST 2002 Speaker Recognition Evaluations (SRE) (non-ideal recordings) [17], [18]. Other systems based on Support Vector Machines (SVMs) using Gaussian supervectors have been proposed and in some cases can lead to lower EERs [19], [20].

Until recently, developing a TTS synthesizer for a targeted speaker required a large amount of speech data from a carefully prepared transcript in order to construct the speech model. However, with a state-of-the-art HMM-based TTS

This work was presented in part at the 2010 and 2011 Int. Conf. Acoust. Speech, and Signal Proc. (ICASSP) and at the 2010 Odyssey Speaker Recognition Workshop.

P. L. De Leon is with Klipsch School of Electrical and Computer Engineering, New Mexico State University (NMSU), Las Cruces NM 88003 USA. e-mail: pdeleon@nmsu.edu

M. Pucher is with Telecommunications Research Center Vienna (ftw), 1220 Vienna, Austria. e-mail: pucher@ftw.at

J. Yamagishi is with the University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom. e-mail: jyamagis@inf.ed.ac.uk

I. Hernaez and I. Saratxaga are with University of the Basque Country, Bilbao, Spain 48013. e-mail: {inma, ibon}@aholab.ehu.es

synthesizer [21], the speech model can now be adapted from an average model (derived from other speakers) or a background model (derived from one speaker) using only a small amount of speech data. Moreover, recent experiments with HMM-based speech synthesis systems have also demonstrated that the speaker-adaptive HMM-based speech synthesis is robust to non-ideal speech data that are recorded under various conditions and with varying microphones, that are not perfectly clean, and/or that lack phonetic balance [22], [23]. In [23] a high-quality voice was built from audio collected off of the Internet. This data was not recorded in a studio, had a small amount of background noise, and the microphones varied in the data. Further [24], [25] reported construction of thousands of voices for HMM-based speech synthesis based on corpora such as the Wall Street Journal (WSJ0, WSJ1, and WSJCAM0), Resource Management, Globalphone and SPEECON. Taken together, these state-of-the-art speech synthesizers pose challenges to SV systems.

In prior work, we utilized a state-of-the-art TTS synthesizer and revisited the problem of imposture using a GMM-UBM SV system with a small speech corpus [26] and then extended to a larger corpus [27]. Recently, we examined the performance using the SVM-based SV system and initial experiments on detecting a synthetic speech signal [28]. In this paper, we provide complete evaluations using both GMM-UBM and SVM-based SV systems and results from a proposed synthetic speech detector which uses phase-based features for classification. First, we train two different SV systems (GMM-UBM and SVM using Gaussian supervectors) using human speech (283 speakers from the WSJ corpus). Second, we create synthetic test speech for each of the 283 speakers by adapting a background model to the targeted speaker. Finally, we measure FAR/FRR when tested using human speech and measure the matched claim rate using synthetic speech. As we will demonstrate, the matched claim rate is above 90% for each of the SV systems hence the vulnerability of the SV systems to synthetic speech. Next, we turn our attention to detection of synthetic speech as a means to prevent imposture by synthetic speech. We summarize results with a previously-proposed method which uses average inter-frame difference of log-likelihood (IFDLL) and show that this is no longer a viable discriminator for high-quality synthetic speech such as that which we are using. Instead, we propose a new discrimination feature based on relative phase shift (RPS) and show that this can be used to reliably detect synthetic speech. We also show a simple and effective method for training the classifier using transcoded human speech as a surrogate for synthetic speech.

This paper is organized as follows. In Sections II and III, we provide overviews of the SV and TTS systems. In Section IV, we review IFDLL and provide details on our proposed RPS feature for detecting synthetic speech. In Section V, we describe the WSJ corpus and explain how we partitioned the corpus for training and testing of all the required systems. We note that although the WSJ journal corpus is not a standard corpus for SV research, it is one of the few that provides sufficient speech material from hundreds of speakers which is required to construct synthetic voices matched to their human counterparts. Section VI gives the evaluation results using the WSJ corpus and its synthesized counterpart as well as the results when using RPS to detect synthetic speech. Finally, we conclude the article in Section VII.

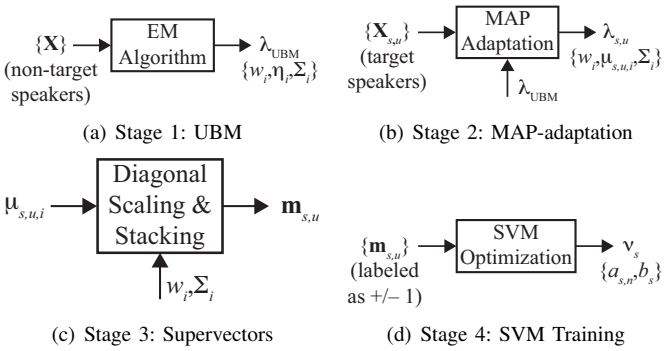


Fig. 1. Stages of training the SV systems. The GMM-UBM SV system is trained with (a)-(b) and the SVM SV system is trained with (a)-(d). Although the GMM-UBM is normally derived from non-target speakers, as described in Section V, we have used target speakers.

II. Speaker Verification Systems

Our SV systems are based on the well-known GMM-UBM described in [17] and the SVM using Gaussian supervectors described in [19]. We briefly review these systems and our implementation in the following subsections.

A. SV System Training

For both SV systems, feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are extracted every 10 ms using a 25 ms hamming window and composed of 15 MFCCs, 15 delta MFCCs, log energy, and delta-log energy as elements.

Training the GMM-UBM system is composed of two stages, shown in Fig. 1(a) and (b). The SVM using Gaussian supervectors system includes these two stages and two additional stages shown in Fig. 1(c) and (d). In the first stage, a GMM-UBM consisting of the model parameters $\lambda_{\text{UBM}} = \{w_i, \eta_i, \Sigma_i\}$ is constructed from the collection of speakers' feature vectors. Here, we assume M component densities in the GMM-UBM and w_i, η_i, and Σ_i represent respectively the weight, mean vector, and diagonal covariance matrix of the i-th component density where $1 \leq i \leq M$. These parameters are estimated using the Expectation Maximization (EM) algorithm. In practice the GMM-UBM is constructed from non-target speakers.

In the second stage, feature vectors are extracted from target speakers' utterances. We assume the availability of several utterances per speaker recorded (preferably) under different channel conditions in order to improve the speaker modeling and robustness of the system. Feature vectors from each utterance are used to maximum a posteriori (MAP)-adapt only the mean vectors of the GMM-UBM to form speaker- and utterenace-dependent models $\lambda_{s,u} = \{w_i, \mu_{s,u,i}, \Sigma_i\}$ where $\mu_{s,u,i}$ is the MAP-adapted mean vector of the i-th component density from speaker s and utterance u.

In the third stage (used for the SVM), the mean vectors $\mu_{s,u,i}$ are then diagonally-scaled according to

$$\mathbf{m}_{s,u,i} = \sqrt{w_i}\Sigma_i^{-1/2}\mu_{s,u,i} \quad (1)$$

and stacked to form a Gaussian supervector for a speaker's given utterance

$$\mathbf{m}_{s,u} = \begin{bmatrix} \mathbf{m}_{s,u,1} \\ \vdots \\ \mathbf{m}_{s,u,N} \end{bmatrix}. \quad (2)$$

In the fourth stage (used for the SVM), the target speaker's supervectors are labeled as $+1$ and all other speakers' supervectors as -1. Parameters (weights, a_n and bias, b) of the SVM using a linear kernel are computed for each speaker through an optimization process. As derived in [29], an appropriately-chosen distance measure between the mean vectors $\mu_{s,u,i}$, results in a corresponding linear kernel involving the supervectors in (2) composed of diagonally-scaled mean vectors (1).

In conventional GMM-UBM SV systems, we normally assume a single training signal (or several utterances concatenated to form a single training signal) so that the speaker model is simply $\lambda_s = \{w_i, \mu_{s,i}, \Sigma_i\}$. For the SVM, the speaker model is denoted $\nu_s = \{a_{s,n}, b_s\}$ where $1 \leq n \leq N$ and N is the total number of supervectors.

B. SV System Testing

In SV system testing we are given an identity claim C and feature vectors $\mathbf{X}$ from a test utterance and must accept or reject the claim. For the GMM-UBM system, we compute the log-likelihood ratio

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_C) - \log p(\mathbf{X}|\lambda_{\mathrm{UBM}}). \quad (3)$$

where

$$\log p(\mathbf{X}|\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{x}_n|\lambda) \quad (4)$$

and N is the number of test feature vectors. The claimant speaker is accepted if

$$\Lambda(\mathbf{X}) \geq \theta \quad (5)$$

where θ is the decision threshold. In the SVM system, the supervector $\mathbf{m}_{\mathrm{test}}$ is computed from the feature vectors $\mathbf{X}$ by essentially repeating stages 2 and 3 from training. We then compute

$$y(\mathbf{X}) = \sum_{n \in \mathcal{S}} a_{C,n} t_{C,n} \mathbf{m}_{\mathrm{test}}^T \mathbf{m}_{C,n} + b_C \quad (6)$$

and accept the claim if $y(\mathbf{X}) \geq 0$. We denote $\mathcal{S}$ as the set of indexes of the support vectors and $t_{C,n}$ as the labels associated with the supervectors.

III. Text-to-Speech Synthesizer

Our TTS system is built using the framework from the "HTS-2008" system [22], [30], which was a speaker-adaptive system entered for the Blizzard Challenge 2007 [31] and 2008 [32]. In the 2008 challenge, the system had the equal best naturalness and the equal best intelligibility on a training data set comprising one hour of speech. The system was also found to be as intelligible as human speech [30]. The speech synthesis system, outlined in Fig. 2, consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

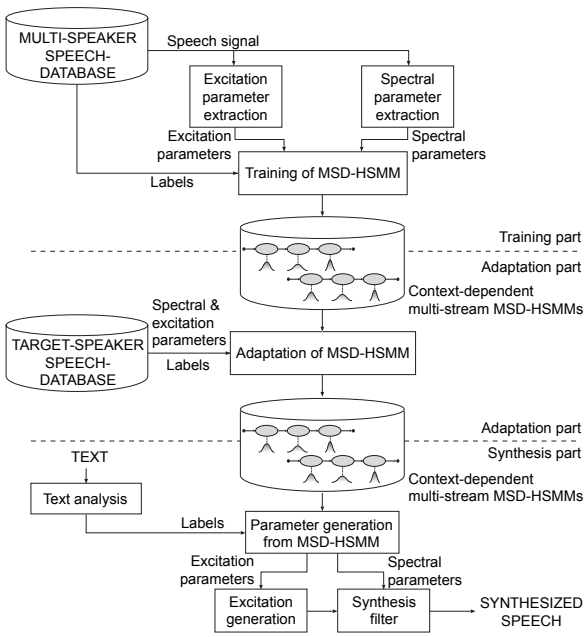


Fig. 2. Overview of the HTS-2008 speech synthesis system, which consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis component, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [33]) mel-cepstral vocoder with mixed excitation (i.e., the mel-cepstrum, $\log F_0$ and a set of band-limited aperiodicity measures) are extracted as feature vectors for HMMs [34]. In the average voice training part, context-dependent, multi-stream, left-to-right, multi-space distribution (MSD), hidden semi-Markov models (HSMMs) [35] are trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. A set of model parameters (Gaussian mean vectors and diagonal covariance matrices) for the speaker-independent MSD-HSMMs are estimated using the EM algorithm.

The training stages for the average voice models are shown in Fig. 3. First, speaker-independent monophone MSD-HSMMs are trained from an initial segmentation, converted into context-dependent MSD-HSMMs, and re-estimated. Then, decision-tree-based context clustering with the MDL criterion [36] is applied to the HSMMs and the model parameters of the HSMMs are tied at leaf nodes. The clustered HSMMs are re-estimated again. The clustering processes are repeated twice and the whole process is further repeated twice using segmentation labels refined with the trained models in a bootstrap manner. All re-estimation and re-segmentation processes utilize speaker-adaptive training (SAT) [37] based on constrained maximum likelihood linear regression (CMLLR) [38].

In the speaker adaptation component the speaker-

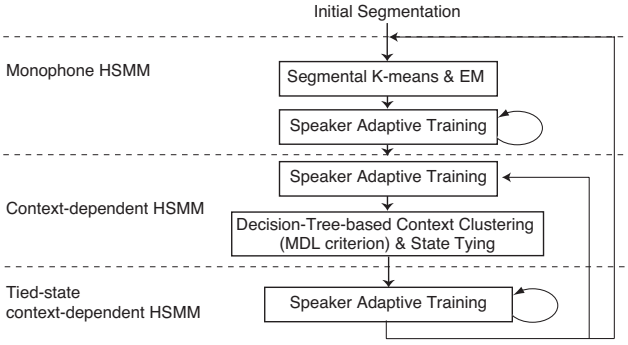


Fig. 3. Overview of the training stages for average voice models.

independent MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression (CSMAPLR) [39]. Note that not only output pdfs for the acoustic features but also duration models are also transformed in the speaker adaptation [40]. In the speech generation component, acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of a trajectory to be generated and trajectory likelihood [41]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [42]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter [43] corresponding to the STRAIGHT mel-cepstral coefficients to generate the synthetic speech waveform.

IV. Detection of Synthetic Speech

In this section, we investigate the problem of detection of synthetic speech. We begin with a method which uses the average IFDLL as proposed in [44]. We then propose a new discrimination feature based on RPS and develop a classifier.

A. Average inter-frame difference of log-likelihood

The IFDLL is defined as [44]

$$\Delta_n \quad = \quad |\log p(\mathbf{x}_n|\lambda_C) - \log p(\mathbf{x}_{n-1}|\lambda_C)| \qquad (7)$$

and the average IFDLL is given by

$$\bar{\Delta} \quad = \quad \frac{1}{N}\sum_{n=1}^{N}\Delta_n. \qquad (8)$$

The authors in [44] observed that for synthetic speech, average IFDLL is significantly lower than that for human speech and can be used as a discriminator. This difference was explained as a result of the HMM-based synthesizer, used in the work, generating a speech parameter sequence so as to maximize the output probability. This maximization normally leads to a time variation of the speech parameters of synthetic speech becoming smaller than that for human speech.

In Fig. 4 we show the distributions of average IFDLL for human and synthetic speech using the 283 speaker WSJ corpus (subsets HS-B and TTS-B as described in Section IV). Using the state-of-the-art HMM-based speech synthesizer described

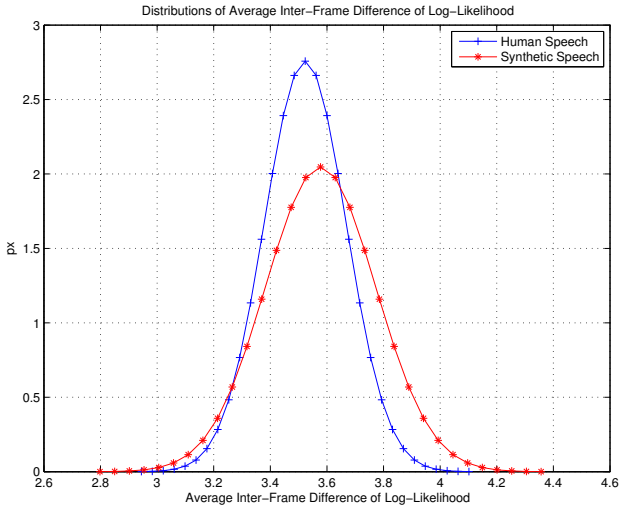


Fig. 4. Distributions of average interframe-difference of log-likelihood for human and synthetic speech. Due to the overlapping distributions, the average IFDLL cannot be used to detect synthetic speech.

in Section III, this measure no longer appears to be robust enough to detect synthetic speech, since the distributions in average IFDLL for human and synthetic speech have significant overlap. In [27] we showed that IFDLL, dynamic-time-warping of MFCC features, and automatic speech recognition word-error-rate are not robust measures to detect synthetic speech. The similar average IFDLL distributions can be explained because the state-of-the-art HMM-based speech synthesizer that we use, includes global time variation models [41].

B. Relative Phase Shift

Since the human auditory system is known to be relatively insensitive to the speech signal's phase [45], the vocoder used in TTS is normally based on a minimum-phase vocal tract model for simplicity. This simplification leads to differences in the phase spectra between human and synthetic speech which are not usually audible. However, these differences can be used to construct a feature which allows detection of synthetic speech.

We propose using the RPS representation of the harmonic phase, which is a simple representation of signal phase for harmonic speech models, as a discriminating feature for detecting synthetic speech [46], [47]. RPS can be defined as follows. The harmonic part of the speech signal may be represented as

$$h(t) = \sum_k A_k(t)\cos\left(\Phi_k\left(t\right)\right) \qquad (9)$$

where $A_k(t)$ is the amplitude and

$$\Phi_k(t) = 2\pi f_0 kt + \theta_k \qquad (10)$$

is the instantaneous phase of the k-th harmonic. Here we denote the fundamental frequency as f_0 and initial phase of the k-th harmonic as θ_k. The RPS values for every harmonic are then calculated from the instantaneous phase $\Phi_k(t)$ at each analysis instant t_a using

$$\text{RPS}_k = \Phi_k(t_a) - k\Phi_1(t_a). \qquad (11)$$

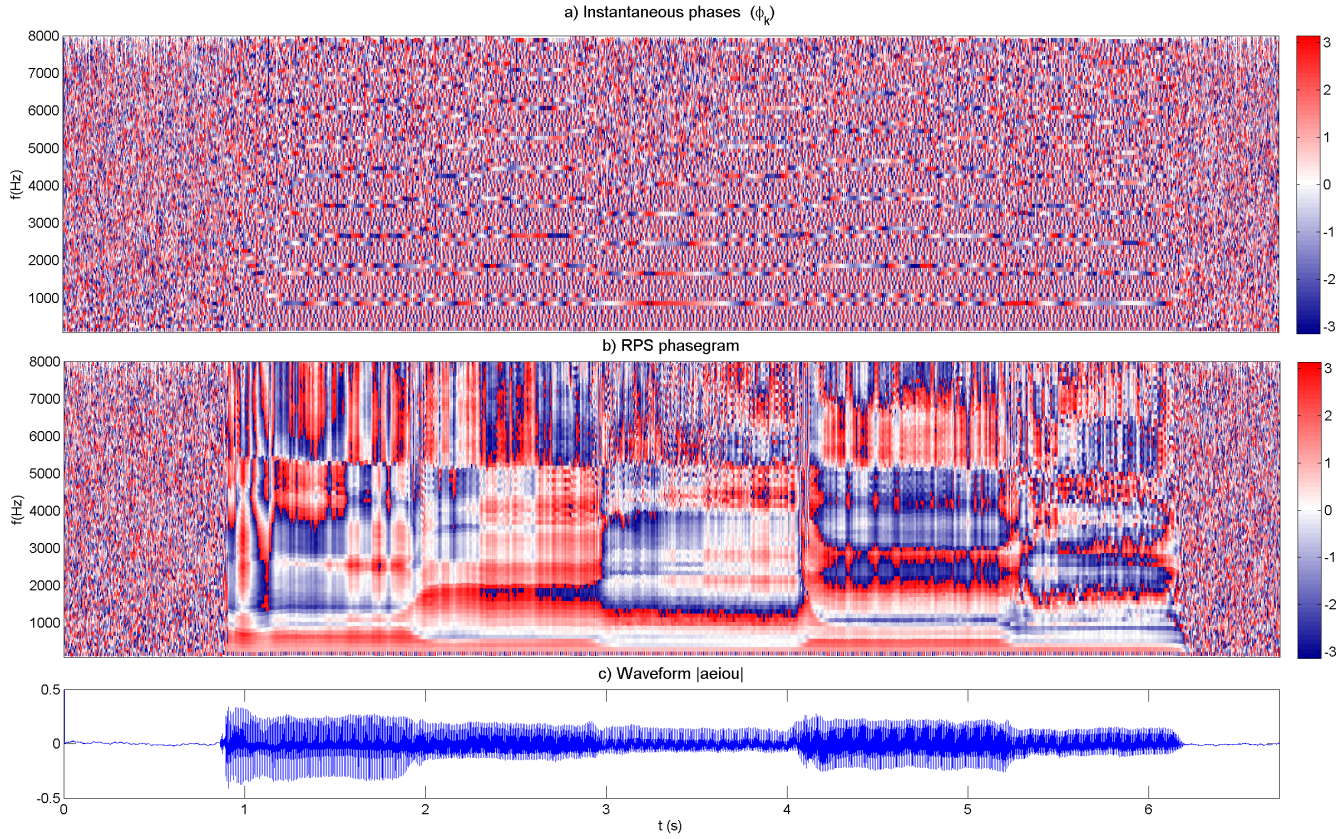


Fig. 5. Phasegrams of a voiced speech segment for five continuous vowels. a) Intantaneous phases b) Relative phase shift c) Signal waveform

More specifically, this transformation removes the linear phase contribution due to the frequency of every harmonic from the instantaneous phase and allows a clear phase structure to arise, as shown in Fig. 5. The RPS values for voiced segments are illustrated in Fig. 5(b) and show a structured pattern along frequency as the signal evolves.

In order to use RPS values as features for classification and detection of synthetic speech, several important steps must be carried out. These steps were initially developed for an Automatic Speech Recognition (ASR) task [47] and are listed below: [47]

1) Due to the variable number of harmonics found in a predefined frequency range, the dimensionality of the vector of RPS values varies from frame to frame. We transform the variable-dimension vectors into fixed-dimension vectors by applying a Mel-scale filter bank with a constant number of filters.
2) The dimensionality of the RPS vector is very high, if the usual analysis bandwidth is considered. This is problematic for training any statistical model, therefore RPS values are computed over a 4 kHz bandwidth and the Discrete Cosine Transform (DCT) is used at the end of the process to decorrelate and reduce the dimensionality.
3) The RPS values in (11) are wrapped phase values and therefore may create discontinuities as shown in Fig. 6(a)-(b). This is also problematic for parametrization. Therefore we unwrap the phase in order to avoid the discontinuities in the RPS envelope.
4) The unwrapping process is ambiguous and very different results may be obtained with similar data as shown in Fig. 6(c)-(d). Therefore we differentiate the unwrapped RPS values in order to alleviate the ambiguity problem as illustrated in Fig. 6(e)-(f).

In order to develop a classifier for synthetic speech, we compute 20 coefficients per speech frame according to steps 1-4. The mean of the differentiated unwrapped RPS (i.e. the mean slope of the unwrapped RPS) has been removed before calculating the DCT and added as a parameter, resulting in a total of 21 coefficients per frame which are used as a feature vector, $\mathbf{y}_t$ for the classifier. Here only voiced segments of the signals have been used, because there is no useful phase information in unvoiced frames. The RPS values are then extracted using a 10 ms frame-rate. Fig. 7 shows a spectrogram-like representation of the parameters (i.e. time in horizontal axis, frequency in the vertical axis and parameter value in grey or coloured scale) obtained for one of the speakers, both for the human speech and his synthetic counterpart. We can see clear differences between the human and synthetic speech in the figure. We use a 32-component density GMM in the classifier trained on RPS feature vectors extracted from human and synthetic speech signals.

Detection of synthetic speech occurs once the speaker verification system has accepted the identity (see Fig. 8)–currently, we see no need to apply the synthetic speech detector (SSD) if the SV system has rejected the identity. If an identity claim,

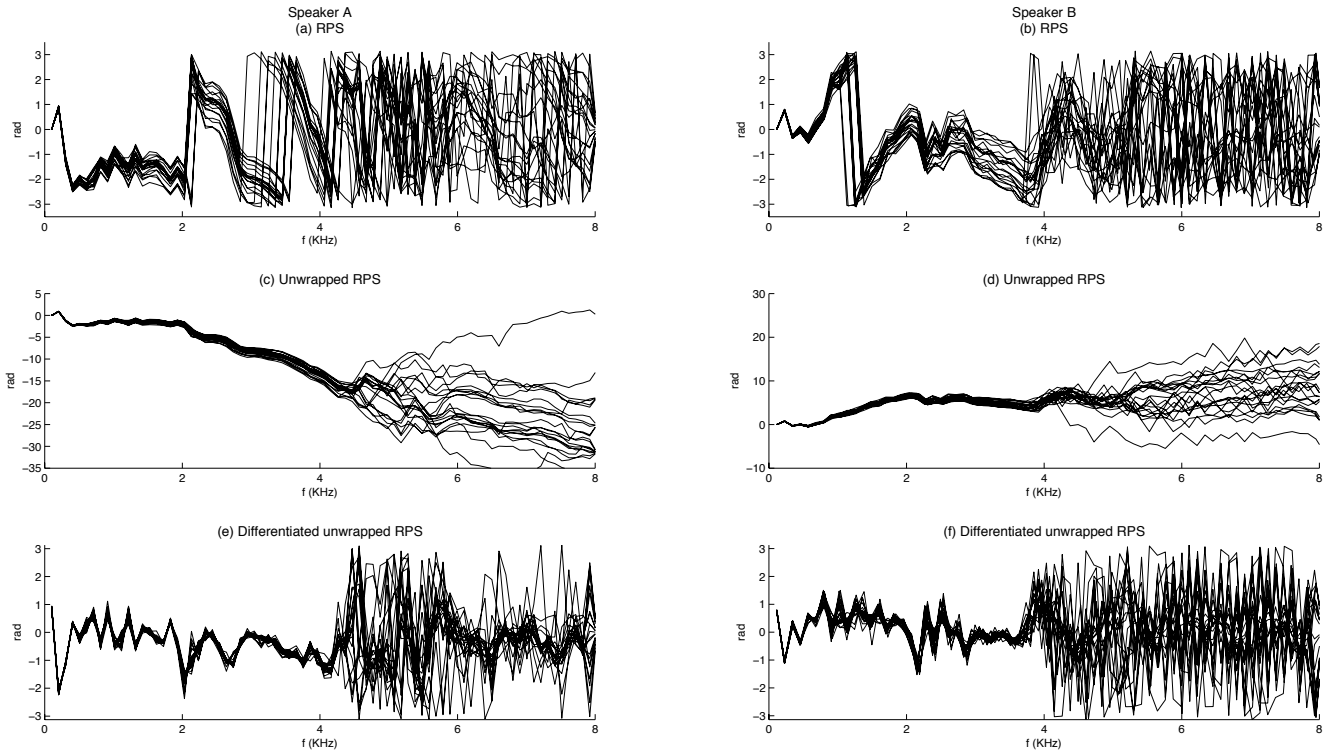


Fig. 6. RPS information for two sustained —i— speech segments of 200ms (20 frames) by two male speakers: (a-b) RPS, (c-d) unwrapped RPS, (e-f) differentiation of the unwrapped RPS

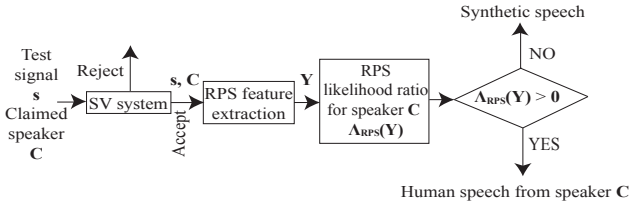


Fig. 8. Proposed system for detection of synthesized speech after speaker verification using phase-based detection.

C is accepted, we compute the log-likelihood ratio

$$\Lambda_{\text{RPS}}(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{C,\text{human}}) - \log p(\mathbf{Y}|\lambda_{C,\text{synth}}) \quad (12)$$

where $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ is the sequence of RPS feature vectors and $\lambda_{C,\text{human}}$ and $\lambda_{C,\text{synth}}$ represent GMMs of the feature vectors for human speech and synthetic speech associated with claimant C, respectively. The input test signal is then classified as human speech if $\Lambda_{\text{RPS}}(\mathbf{Y}) > 0$, otherwise it is classified as synthetic.

V. Data Sets

For this research, we use the WSJ corpus from the Linguistic Data Consortium (LDC) [48]. Although the WSJ journal 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 [49]. From the corpus, we chose the pre-defined official training data set, SI-284, that includes both WSJ0 and WSJ1 as material data. The SI-284 set has a total of 81 hours of speech data uttered by 283 speakers[1] and was partitioned into three disjoint "human speech" subsets HS-A, HS-B, and HS-C, as shown in Table I. Subset HS-A was used to train the TTS system described in Section III, subset HS-B was used to train the SV and SSD systems described in Sections II and VI-B, and subset HS-C was used to test the SV and SSD systems. Once trained, the TTS system was used to generate the synthetic speech subsets TTS-B and TTS-C as shown in Table I which are used to train the SSD and test the SV/SSD systems respectively. These different subsets were used to avoid any overlapping of data sets and associated cross-corpus negative effects while attempting to simulate realistic imposture scenarios[2].

Training the SSD with synthetic speech has a practical disadvantage, that is, a TTS synthesizer has to be trained for each speaker in the SV system. Therefore, we have also evaluated a more practical method that uses the STRAIGHT vocoder to transcode the human speech signal as a surrogate for TTS-generated (synthesized) speech. By transcoding, the human speech signal is parametrized using a vocoder and from this parametrization, the speech signal is reconstructed in a process similar to that in the TTS speech generation component. The transcoded human speech signal has artifacts similar to those in the synthetic speech signal which can be useful for simplifying the training of the SSD. In order to

[1]One speaker was removed from the data due to poor recording conditions.

[2]In future work, the average voice model of the TTS should be derived from a different corpus.

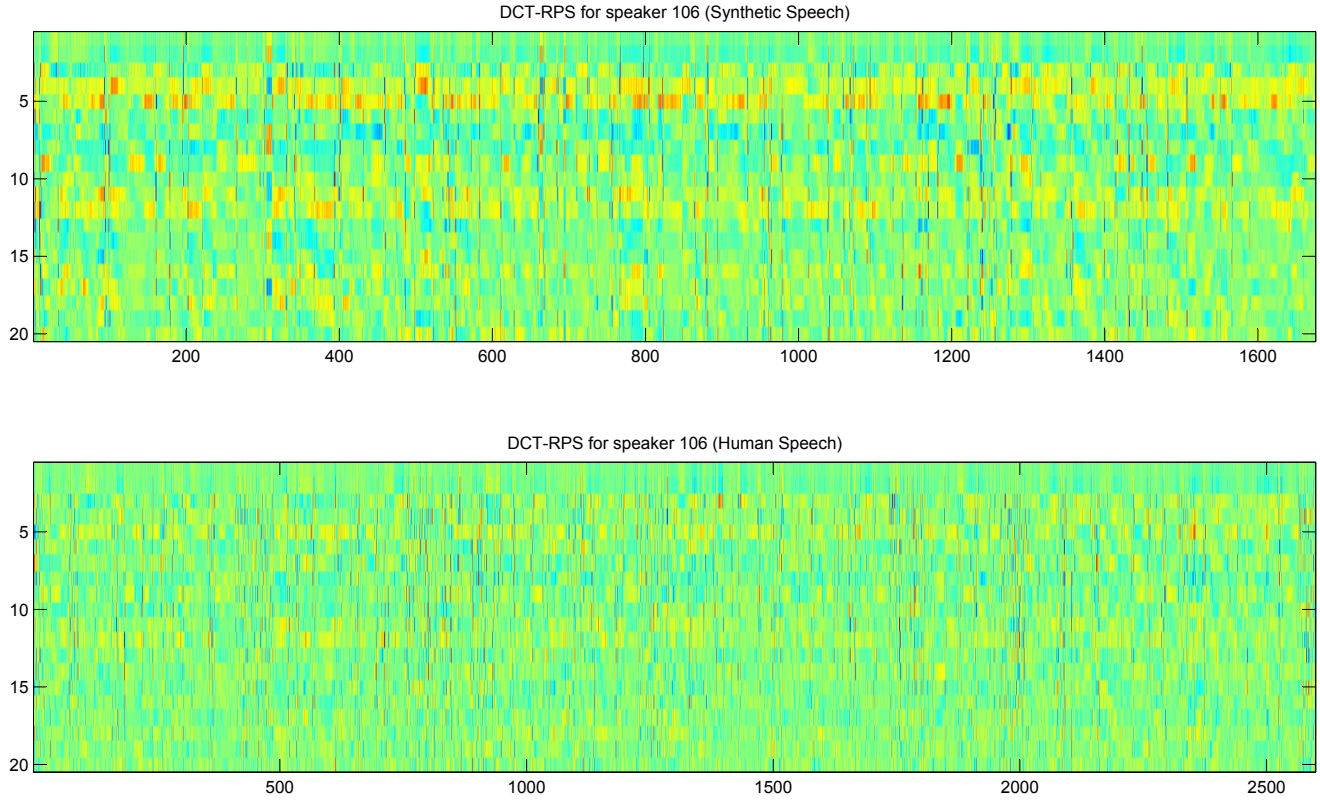


Fig. 7. Spectrogram-like representation of the parametrized RPS for a real signal (upper) and a synthesized signal (below).

TABLE I
WALL STREET JOURNAL (WSJ) CORPUS PARTITIONS USED FOR TRAINING AND TESTING OF TEXT-TO-SPEECH (TTS), SPEAKER VERIFICATION (SV), AND SYNTHETIC SPEECH DETECTOR (SSD) SYSTEMS.

Human speech (HS)	HS-A train TTS	HS-B train SV train SSD	HS-C test SV test SSD
Synthetic speech (TTS)		TTS-B train SSD	TTS-C test SV test SSD
transCoded speech (CS)		CS-B train SSD	

evaluate this approach, we transcoded subset HS-B and created the CS-B "coded speech" subset as shown in Table I. By using CS-B instead of TTS-B to train the SSD, all system components (TTS, SV, SSD) can be trained using only human speech.

Since each speaker included in the SI-284 set has different speech durations, we used varying lengths (73 sec to 27 min) of training signals from subset HS-A to construct and adapt the TTS system to each speaker. Some speakers have larger amounts of data than those we can practically collect for the imposture against the SV system.

TABLE II
SPEAKER VERIFICATION RESULTS FOR THE GMM-UBM SYSTEM AND THE SVM USING GAUSSIAN SUPERVECTORS SYSTEM.

	GMM-UBM	SVM
EER (human speech)	0.35%	0.35%
Accepted matched claims (synthetic speech)	259/283 = 91.5%	271/283 = 95.8%

VI. EXPERIMENTS AND RESULTS

A. Speaker Verification

For the two SV systems, we have trained on ≈90s speech signals from subset HS-B and tested using ≈30s signals from subsets HS-C and TTS-C. Training signals for the SVM SV system were segmented into eight utterances per speaker and used to construct Gaussian supervectors as described in Section II-A. The evaluation for human speech was designed so that each test utterance has an associated true claim and 282 false claims yielding 283^2 tests. Test results for each system under human speech are given in row 2 of Table II and the Detection Error Tradeoff (DET) curves are shown in Fig. 9. The low EERs (0.35% for both SV systems) are due to the ideal nature of the recordings in the WSJ corpus and the accuracy of the SV systems. We note that both the GMM-UBM and SVM systems have about the same performance under human speech.

The evaluation for synthetic speech was designed so that

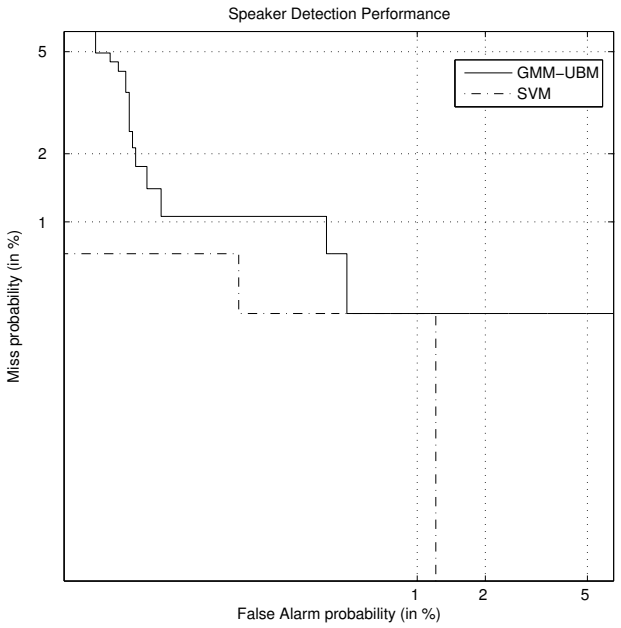
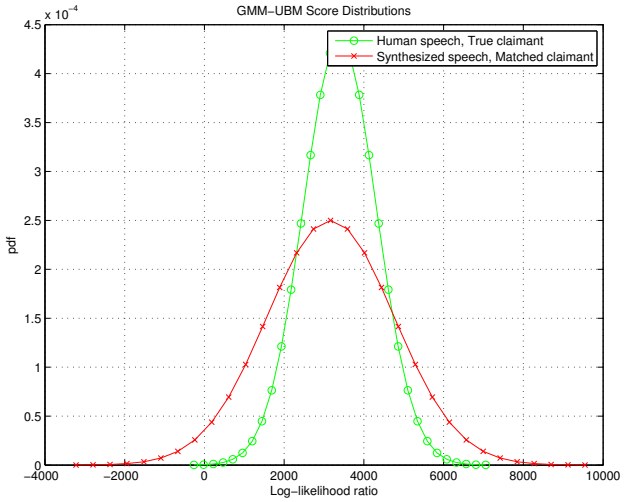
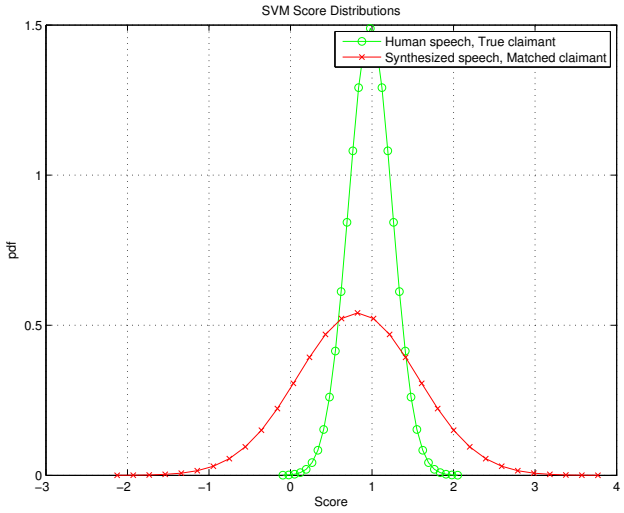


Fig. 9. DET curves for speaker verification using test signals from human speakers. The EER is 0.35% for GMM-UBM and SVM systems.



(a) GMM-UBM SV System



(b) SVM using Gaussian supervectors SV system

Fig. 10. Approximate score distributions for (a) GMM-UBM and (b) SVM using Gaussian supervectors SV systems with human and synthesized speech. Distributions for synthesized speech (red lines) have significant overlap with those for human speech (red lines) leading to over a 91% acceptance rate for synthetic speech with matched claims.

each test utterance has an associated matched claim yielding 283 tests for imposture. (In a realistic imposture scenario, a speech signal targeted at a specific speaker will be synthesized and a claim only for that speaker will be submitted, i.e. matched claim.) For both SV systems, the decision thresholds are chosen for EER under human speech signal tests. Row 3 of Table II shows the results in which we can see over 91% of synthetic speech signals with an associated matched claim will be accepted by the systems. It is interesting to note that the SVM using Gaussian supervectors accepts more claims using synthetic speech than the GMM-UBM despite both systems (under human speech) having the same EER and the SVM system performing slightly better on the DET. As described in an earlier paper, this result is due to significant overlap in the score distributions for human and synthetic speech, as shown in Fig. 10 [26]. Thus, adjustments in decision thresholding or standard score normalization techniques cannot differentiate between true and matched claims originating from human and synthesized speech [50], [51].

B. RPS-Based Detection of Synthetic Speech

We trained the SSD on human speech using HS-B and synthetic speech using TTS-B as in Table I and evaluated classifier accuracy with human speech from HS-C and synthetic speech from TTS-C. These results are shown in row 2 of Table III where we find 100% accuracy in classifying a speech signal as either human or synthetic. We also trained the SSD with transcoded speech using CS-B as a surrogate for synthetic speech and set the decision threshold to either zero or 1.65 for EER. These results are shown in row 3 of Table III where we find with the decision threshold set to zero, human speech signals are classified with 100% accuracy and synthetic speech signals are classified with 90.10% accuracy. With the decision threshold set to 1.65 for EER, we find 97.17% accuracy in classifying a speech signal as either human or synthetic. Approximate distributions for the classifier scores, $\Lambda_{\text{RPS}}(\mathbf{Y})$ are shown in Fig. 11 where we see with transcoded speech (CS-B models) it is necessary to adjust the decision threshold slightly upward for EER.

Next, we evaluated the overall system which includes the SSD and SV systems as illustrated in Fig. 8. Table IV shows acceptance rates for human speech for true claimants and acceptance rates for synthetic speech for matched claims for the overall system. Both GMM-UBM and SVM SV systems are considered, with and without the SSD as illustrated in Fig. 8. For convenience, the first row repeats the earlier results (no synthetic speech detection) from Table II illustrating the problem. Using the proposed SSD trained on TTS-B, the acceptance rate for synthetic speech is now reduced from over 91% to 0% with no change in the acceptance rate for

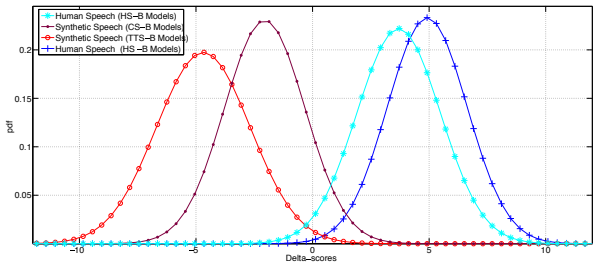


Fig. 11. Approximate distributions for the classifier scores, $\Lambda_{\mathrm{RPS}}(\mathbf{Y})$ when tested with human and synthetic speech. Blue and red curves show classifier performance when trained on human speech using HS-B and synthetic speech TTS-B. Cyan and majenta curves show classifier performance when trained on human speech using HS-B and transcoded speech CS-B. Both classifiers were tested with human speech using HS-C and synthetic speech using TTS-C.

TABLE III
ACCURACY RATES FOR THE CLASSIFICATION OF HUMAN AND SYNTHETIC SPEECH. CLASSIFIER IS TRAINED WITH HUMAN SPEECH SUBSET HS-B AND EITHER TTS-B OR CS-B FOR SYNTHETIC SPEECH. CLASSIFIER IS TESTED USING SUBSETS HS-C AND TTS-C. RESULTS ARE BASED ON A ZERO THRESHOLD FOR LOG-LIKELIHOOD RATIO (12) AND INCLUDE AN ADDITIONAL RESULT FOR CS-B WHERE THRESHOLD IS ADJUSTED FOR EER.

	Accuracy rate of classifier	
Training Data	Human Speech (HS-C)	Synthetic Speech (TTS-C)
HS-B/TTS-B	100%	100%
HS-B/CS-B	100%, 97.17%	90.10%, 97.17%

human speech. As mentioned earlier, constructing synthetic voices for each human registered in the SV system is not very practical, so we proposed training the SSD using transcoded human speech as a surrogate for synthetic speech. Training the SSD on CS-B, results in an acceptance rate for synthetic speech of 9.5%, 9.9% for the GMM-UBM, SVM SV systems, respectively with no change in the acceptance rate for human speech. Finally, adjusting the decision threshold in the SSD for EER, we can reduce acceptance rate for synthetic speech to below 2.8% with a slight decrease in acceptance rate for human speech (from 99.7% to 96.8%). From these results, we conclude that the SSD trained on transcoded speech can drastically reduce the number of accepted matched claims associated with synthetic speech, while maintaining SV accuracy for human speech. Thus the proposed method is an accurate and effective method for securing the SV systems against the imposture using synthetic speech.

VII. CONCLUSIONS

In this paper, we have evaluated the vulnerability of speaker verification (SV) to imposture using synthetic speech. Using the Wall Street Journal corpus and two different SV systems (GMM-UBM and SVM using Gaussian supervectors), we have shown that with state-of-the-art speech synthesis, over 91% of matched claims, i.e. a synthetic speech signal matched to a targeted speaker and an identity claim of that same speaker, are accepted. Thus despite the excellent performance of the SV systems under human speech, the quality of synthesized speech is high enough to allow these synthesized

TABLE IV
ACCEPTANCE RATES FOR HUMAN SPEECH (TRUE CLAIMANT) AND SYNTHETIC SPEECH (MATCHED CLAIM) FOR OVERALL SYSTEM CONSISTING OF SPEAKER VERIFICATION AND SYNTHETIC SPEECH DETECTOR (SSD). IDEALLY THE SYSTEM HAS 100% ACCEPTANCE RATE FOR HUMAN SPEECH, TRUE CLAIM AND 0% FOR SYNTHETIC SPEECH, MATCHED CLAIM.

	GMM-UBM	SVM
Without SSD		
Acceptance rate for human, true claim	99.7%	99.7%
Acceptance rate for synthetic, matched claim	91.5%	95.8%
With SSD trained on TTS-B		
Acceptance rate for human, true claim	99.7%	99.7%
Acceptance rate for synthetic, matched claim	0.0%	0.0%
With SSD trained on CS-B		
Acceptance rate for human, true claim	99.7%	99.7%
Acceptance rate for synthetic, matched claim	9.5%	9.9%
With EER SSD trained on CS-B		
Acceptance rate for human, true claim	96.8%	96.8%
Acceptance rate for synthetic, matched claim	2.5%	2.8%

voices to pass for true human claimants. This result suggests that synthetic speech may pose security issues for speech-based remote/online authentication or incorrect speaker confirmation. As a potential solution to the imposture problem, we have proposed a synthetic speech detector (SSD) based on relative phase shift (RPS) features. Although remarkably accurate, training the SSD requires that a TTS synthesizer be constructed for each speaker in the SV system which is not practical. Therefore, we have proposed using transcoded speech as a surrogate for synthetic speech in training the SSD. Our results show that we can reduce the acceptance rate of matched claims using synthetic speech to less than 3%, while maintaining SV accuracy for human speech.

REFERENCES

[1] 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.

[2] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Commun.*, vol. 17, no. 1-2, pp. 109–116, Aug. 1995.

[3] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, Oct. 2004, pp. 145 – 148.

[4] K. Sullivan and J. Pelecanos, "Revisiting carl bildts impostor: Would a speaker verification system foil him?" in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes in Computer Science, J. Bigun and F. Smeraldi, Eds. Springer Berlin / Heidelberg, 2001, vol. 2091, pp. 144–149.

[5] D. Genoud and G. Chollet, "Speech pre-processing against intentional imposture in speaker recognition," in *Proceedings of ICSLP-98, Sidney*, Dec. 1998.

[6] B. L. Pellom and J. H. L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference - Volume 02*, ser. ICASSP '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 837–840.

[7] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification – a study of possible technical imposter techniques." in *Eurospeech-99*, vol. 3, 1999, pp. 1211–1214.

[8] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, p. I.

[9] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech 2007*, April 2007, pp. 2053–2056.

[10] ——, "Transfer function-based voice transformation for speaker recognition," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, June 2006, pp. 1 –6.

[11] M. Farrus, D. Erro, and J. Hern, "Speaker recognition robustness to voice conversion," *IV Jornadas de Reconocimiento Biometrico de Personas*, pp. 73–82, Sept. 2008.

[12] L.-J. Bo, "Forensic voice identification in france," *Speech Communication*, vol. 31, no. 2-3, pp. 205 – 224, 2000.

[13] 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. EUROSPEECH*, 1999.

[14] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP*, 1996.

[15] ——, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP*, 1997.

[16] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. ICSLP*, 2000.

[17] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleaua, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Applied Signal Process.*, vol. 4, pp. 430–451, 2004.

[18] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 277–288, Jan. 2006.

[19] 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.

[20] C. Longworth and M. Gales, "Combining derivative and parametric kernels for speaker verification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 748–757, May 2009.

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

[22] 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 & Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[23] J. Yamagishi, Z.-H. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. Interspeech 2008*, Brisbane, Australia, Sep. 2008, pp. 581–584.

[24] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, UK, September 2009, pp. 420–423.

[25] ——, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio & Language Process*, vol. in press, March 2010.

[26] P. L. D. 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)*, Dallas, USA, March 2010.

[27] P. L. D. 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.

[28] P. L. D. 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)*, Dallas, USA, 2011.

[29] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, pp. 97–100.

[30] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Sep. 2008.

[31] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[32] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, September 2008.

[33] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[34] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[35] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[36] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[37] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[38] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[39] 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, 1 2009.

[40] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[41] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[42] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–468, 1990.

[43] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, Mar. 1992, pp. 137–140.

[44] 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. Eurospeech*, 2001.

[45] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Prentice-Hall, Inc., 2002.

[46] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, pp. 381–383, 2009.

[47] I. Saratxaga, I. Hernaez, I. Odriozola, E. Navas, I. Luengo, and D. Erro, "Using harmonic phase information to improve asr rate," in *Interspeech*, Japan, 2010.

[48] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, New York, 1992, pp. 357–362.

[49] "Wall Street Journal Corpus," 2010. [Online]. Available: {http://www.ldc.upenn.edu}

[50] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," *Proc. IEEE. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 595–598, April 1988.

[51] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for test-independent speaker verification system," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.