Evaluation of Speaker Verification Security and Detection of HMM-based 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 GMM 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 81% of the matched claims are accepted. This result suggests 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 [?]. 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 targeted speaker and the well-known solution is a text-prompted approach [?]. In addition, the vulnerability of SV to voice mimicking by humans has also been examined in [?], [?]. 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 [?], [?], [?], voice conversion of the recorded voice [?], [?], [?], [?], and diphone speech synthesis methods [?].

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 a 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 [?] 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. [?]. In their original work, the authors used a hidden Markov model (HMM)-based text-prompted SV system [?] and an HMM-based TTS synthesizer. In the SV system, feature vectors were scored against speaker and background models composed of concatenated phoneme models. The acoustic models used in the speech synthesizer were adapted to each of the human speakers [?], [?]. 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. [?], the authors extended the research in two ways. First, they improved their synthesizer by generating speech using F_0 (fundamental frequency) information. Second, they improved their SV system by utilizing both F_0 and spectral information. The F_0 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 [?]. Since this time, GMM-UBM based SV systems have produced excellent

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

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. This research was partly funded by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG. The work of J. Yamagishi was supported by the Engineering and Physical Sciences Research Council (EPSRC) grants EP/I031022/1 and EP/J002526/1. This work has been partially supported by the Spanish Gov. (TEC2009-14094-C04-02), the Basque Government (IE09-262, MV20090225).

performance and have achieved equal error rates (EERs) of 0.1% on the TIMIT corpus (ideal recordings) and 12% on NIST 2002 Speaker Recognition Evaluations (SRE) (non-ideal recordings) [?], [?]. Newer systems based on support vector machines (SVMs) using GMM supervectors have been proposed and in some cases can lead to lower EERs [?], [?].

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 synthesizer [?], 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. In [?] 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 [?] 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 new 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 [?] and then extended to a larger corpus [?]. Recently, we examined the performance using the SVM-based SV system and initial experiments on detecting a synthetic speech signal [?]. In this paper, we provide complete evaluations using both GMM-UBM and SVM-based SV systems and new results from a proposed synthetic speech detector (SSD) which uses phase-based features for classification. First, we train two different SV systems (GMM-UBM and SVM using GMM 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 EER and true acceptance rates when tested using human speech and measure the matched claim rate using synthetic speech. As we will demonstrate, the matched claim rate is above 81% 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 brief 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 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 **??**.

II. Speaker Verification Systems

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

A. SV System Training

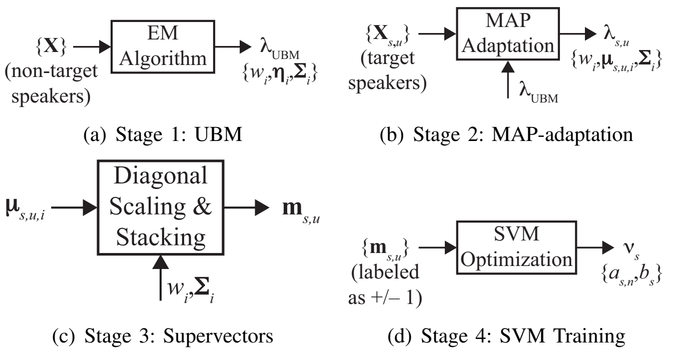
For both SV systems, T feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are extracted every 10 ms using a 25 ms hamming window and composed of 15 mel-frequency cepstral coefficients (MFCCs), 15 delta MFCCs, log energy, and delta-log energy as elements. We apply feature warping to the vectors in order to improve robustness [?] which is adequate given the high-quality recordings in the WSJ corpus.

Training the GMM-UBM system is composed of two stages, shown in Fig. 1(a) and (b). The SVM using GMM 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 $\lambda_{\mathrm{UBM}} = \{w_i, \boldsymbol{\eta_i}, \boldsymbol{\Sigma}_i\}$ is constructed from the collection of speakers' feature vectors. Here, we assume $M = 512$ component densities in the GMM-UBM and w_i, $\boldsymbol{\eta_i}$, and $\boldsymbol{\Sigma}_i$ 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 utterance-dependent models $\lambda_{s,u} = \{w_i, \boldsymbol{\mu}_{s,u,i}, \boldsymbol{\Sigma}_i\}$ where $\boldsymbol{\mu}_{s,u,i}$ 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 $\boldsymbol{\mu}_{s,u,i}$ are then diagonally-scaled according to

$$\mathbf{m}_{s,u,i} = \sqrt{w_i}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_{s,u,i} \qquad (1)$$



(a) Stage 1: UBM

(b) Stage 2: MAP-adaptation

(c) Stage 3: Supervectors

(d) Stage 4: SVM Training

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.

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

$$\mathbf{m}_{s,u} \quad = \quad \begin{bmatrix} \mathbf{m}_{s,u,1} \\ \vdots \\ \mathbf{m}_{s,u,M} \end{bmatrix}. \tag{2}$$

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 [?], an appropriately-chosen distance measure between the mean vectors $\boldsymbol{\mu}_{s,u,i}$, 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 s-th 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 $a_{s,n}$ is the weight of the n-th support vector, b_s is the bias, and $n \in \mathcal{S}$ and $\mathcal{S}$ is the set of indices of the support vectors.

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

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

where

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

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

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

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

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

where $l_{C,n}$ denotes the labels associated with the support vectors and accept the claim if $y(\mathbf{X}) \geq 0$.

III. Text-to-Speech Synthesizer

Our TTS systems are based on the well-known statistical parametric speech synthesis framework described in [?]. The speaker adaptation techniques of the framework allows us to generate a personalized synthetic voice using as little as a few minutes of recorded speech from a target speaker and we use the techniques for building the personalized synthetic voices for hundreds of speakers [1]. In the following subsections, we briefly review our TTS systems and our implementation.

A. TTS System Training

Our TTS system is built using the framework from the "HTS-2008" system [?], which was a speaker-adaptive system entered for the Blizzard Challenge 2008 [?]. In the 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 [?]. The speech synthesis system consists of three main components: speech analysis and average voice training, speaker adaptation, and speech generation.

In the speech analysis and the average voice training component, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [?]) mel-cepstral vocoder with mixed excitation (i.e., 39-dimensional mel-cepstral coefficients, $\log F_0$ and five-dimensional band-limited aperiodicity measures) are extracted as feature vectors for HMMs [?]. Context-dependent, multi-stream, left-to-right, multi-space distribution (MSD), hidden semi-Markov models (HSMMs) [?] 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. 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

[1] We are not considering unit selection and concatenative speech synthesis which is used in some commercial speech synthesizers [?]. Developing the unit selection and concatenation synthesizer for a targeted speaker requires a large amount of speech data, at least one hour, from a carefully prepared transcript. Therefore, we believe this approach is unlikely to be used, in practice, for imposture against SV systems in contrast to HMM-based TTS systems, which requires much smaller amounts of speech. It is possible, however, to use "voice conversion" techniques to change the speaker in the unit selection synthesizer and there are reports [?], [?], [?], [?] of this approach being used for imposture against SV systems. We note that voice conversion approaches use similar vocoders to statistical parametric speech synthesis and we hypothesize that the proposed synthetic speech detection method would also be effective with voice conversions.

clustering with the minimum description length (MDL) criterion [?] 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) [?] based on constrained maximum likelihood linear regression (CMLLR) [?].

B. TTS System Adaptation

In the speaker adaptation component, the speaker-independent MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression (CSMAPLR) [?]. Note that not only output pdfs for the acoustic features but also duration models are transformed in the speaker adaptation. This adaptation requires as little as a few minutes of recorded speech from a target speaker in order to generate a personalized synthetic voice.

C. TTS System Synthesis

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 [?]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [?]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter [?] corresponding to the STRAIGHT mel-cepstral coefficients to generate the synthetic speech waveform.

IV. Detection of Synthetic Speech

In this section, we begin by evaluating the average IFDLL, previously proposed in [?] to detect synthetic speech. As we demonstrate, average IFDLL is no longer a viable discriminator for state-of-the-art HMM-based synthetic speech such as that which we are using. Based on these results, we then propose a more accurate GMM-based classifier based on the RPS feature. The use of a phase-based feature extracted directly from the speech signal is a novel application in the detection of synthetic speech.

A. Average inter-frame difference of log-likelihood

The IFDLL is defined as [?]

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

and the average IFDLL is given by

$$\bar{\Delta} = \frac{1}{R}\sum_{n=1}^{R}\Delta_n. \tag{8}$$

The authors in [?] 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.

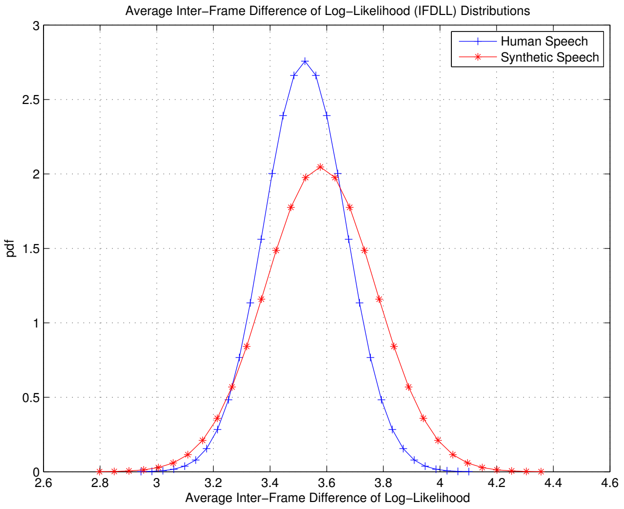


Fig. 2. Approximate distributions of average interframe-difference of log-likelihood for human and synthetic speech. Due to the overlapping distributions, the average IFDLL cannot be used to detect synthetic speech.

In Fig. 2 we show the approximate 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 V). Using the state-of-the-art HMM-based speech synthesizer described in Section III, this measure no longer appears to be robust enough to detect synthetic speech, since the distributions of average IFDLL for human and synthetic speech have significant overlap. In [?], we also showed that dynamic-time-warping of MFCC features and automatic speech recognition (ASR) word-error-rate are also not robust measures to detect synthetic speech.

B. Relative Phase Shift

Since the human auditory system is known to be relatively insensitive to the speech signal's phase [?], 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 new feature which allows detection of synthetic speech.

We propose using the RPS representation of the harmonic phase as a discriminating feature for detecting synthetic speech. The RPS is described in [?], [?] and is based on the harmonic modeling of the speech signal [?]. In these models, the harmonic part of the speech signal may be represented as

$$h(t) = \sum_k A_k(t)\cos\left[\Phi_k\left(t\right)\right] \tag{9}$$

where $A_k(t)$ is the amplitude and

$$\Phi_k(t) = 2\pi F_0 kt + \theta_k \tag{10}$$

is the instantaneous phase of the k-th harmonic. Here we denote the 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

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

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. 3. The RPS values for voiced segments are illustrated in Fig. 3(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 ASR task [?] and are listed below:

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 32 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 frequency range from 0 to 4 kHz 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. 4(a)-(b). This is also problematic for parameterization. Therefore we unwrap the phase in order to avoid the discontinuities in the RPS envelope.
4) Due to its accumulative, non-linear nature, the unwrapping process leads to very different RPS envelopes even if they derive from similar initial data as shown in Fig. 4(c)-(d). If we differentiate the unwrapped RPS envelope the accumulative effect is eliminated, the range of the curve is limited to $[-\pi, \pi]$, and thus similarities between envelopes are more properly perceived. This can be seen in Fig. 4(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 voiced/unvoiced decision is made using the cepstrum-based pitch detection (CDP) algorithm [?]. The RPS values are then extracted using a 10 ms frame-rate.

For the SSD, 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. 5)–currently, we see no need to apply the SSD if the SV system has rejected the identity. If an identity claim, C is accepted, we compute the log-likelihood ratio

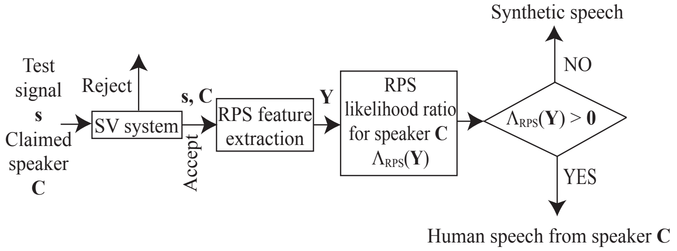


Fig. 5. Proposed system for detection of synthesized speech after speaker verification using phase-based detection.

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

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 RPS feature vectors for human and synthetic speech associated with claimant C, respectively. The speaker is then classified as human 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) [?]. 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 [?]. 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 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 IV-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 and 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 parameterization, the speech signal is reconstructed in a process similar to that in the TTS speech generation

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

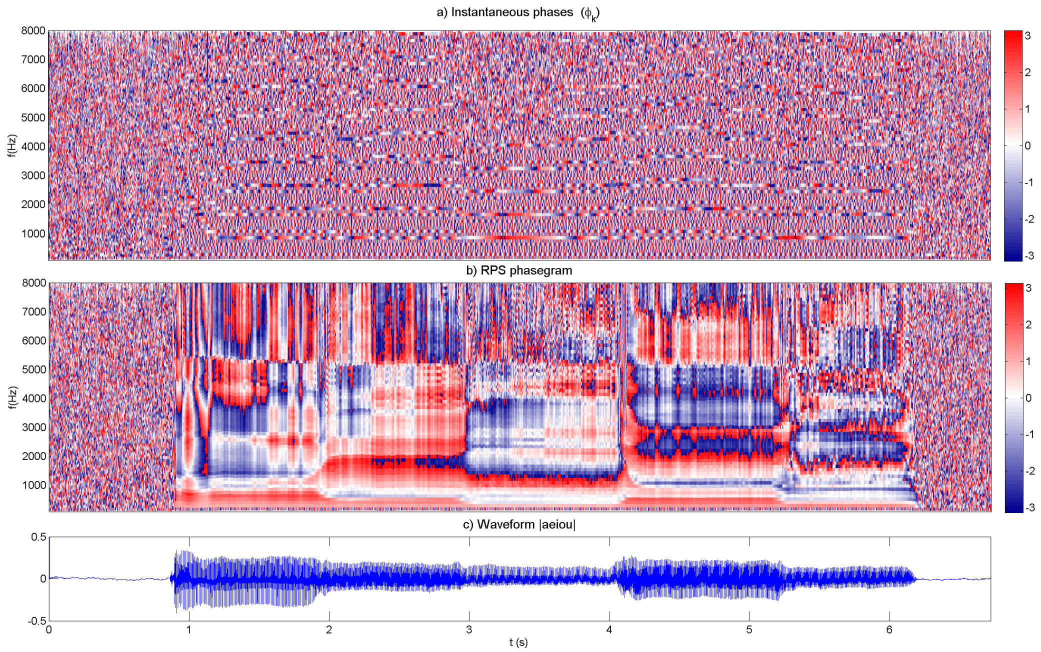


Fig. 3. Phasegrams of a voiced speech segment for five continuous vowels. a) Intantaneous phases b) Relative phase shift c) Signal waveform

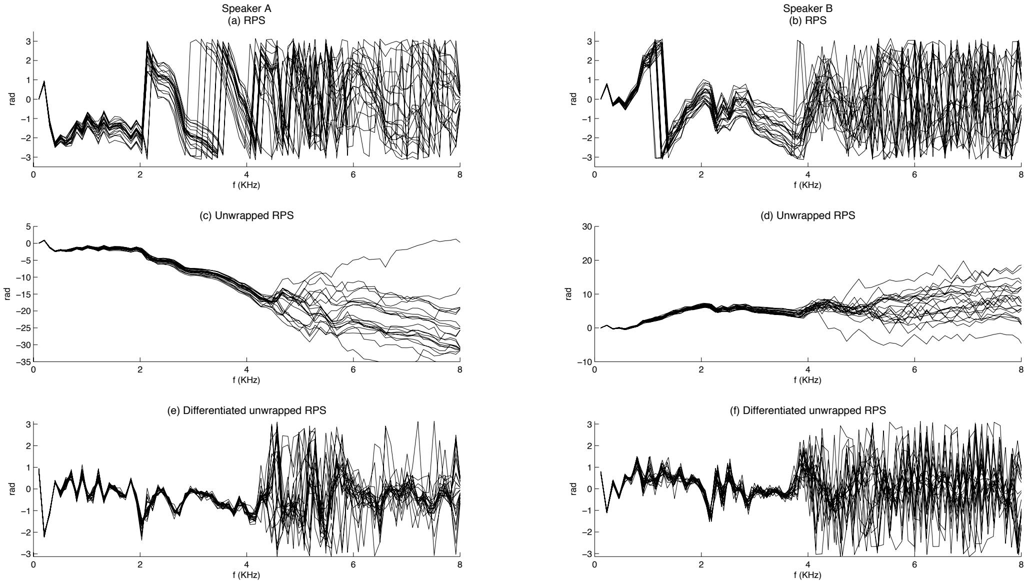


Fig. 4. RPS information for two sustained —i— speech segments of 200ms (20 frames) by two male speakers: (a-b) RPS, (c-d) unwrapped RPS, (e-f) differentiation of the unwrapped RPS

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 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.

VI. Experiments and Results

A. Evaluation of Speaker Verification Systems

For the two SV systems, we have trained using ≈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

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	

used to construct GMM 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 a total of 283^2 tests. The EERs are 0.284%, 0.002% for the GMM-UBM, SVM system respectively. The low EERs ($< 0.3\%$ for both SV systems) are due to the ideal nature of the recordings in the WSJ corpus and the accuracy of the SV systems. Table II row 2 shows the acceptance rates of the SV systems under human speech for true claims as 99.7%, 100% for the GMM-UBM, SVM system respectively.

The evaluation for synthetic speech was designed so that 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. Table II row 3 shows the results where we see over 81% of synthetic speech signals with an associated matched claim will be accepted by the SV systems. 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. 6 [**?**]. Thus, adjustments in decision thresholding or standard score normalization techniques cannot differentiate between true and matched claims originating from human and synthesized speech [**?**], [**?**]. For completeness in Fig. 6, we show the score distributions for synthesized speech, false claim (imposter) even though in the imposture scenario, only matched claims would be submitted.

B. Evaluation of Synthetic Speech Detector

We trained the SSD, described in Section IV-B, 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 Table III row 1 where we find 100% accuracy in classifying a speech signal as either human or synthetic. As mentioned earlier, constructing synthetic voices for each human registered in the SV system is not very practical, so we trained the SSD using transcoded human speech CS-B as a surrogate for synthetic speech. These results are shown in Table III where we find that 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 at 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. **??** where we see that with transcoded speech (CS-B models) it is necessary to adjust the decision threshold slightly upward for EER.

TABLE II
ACCEPTANCE RATES FOR HUMAN SPEECH (TRUE CLAIMANT) AND SYNTHETIC SPEECH (MATCHED CLAIM) FOR OVERALL SYSTEM CONSISTING OF SPEAKER VERIFICATION (SV) 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%	100%
Acceptance rate for synth, matched claim	85.5%	81.3%
With SSD trained on TTS-B		
Acceptance rate for human, true claim	99.6%	100%
Acceptance rate for synth, matched claim	0.0%	0.0%
With SSD trained on CS-B		
Acceptance rate for human, true claim	99.6%	100%
Acceptance rate for synth, matched claim	8.8%	8.8%
With SSD (set for EER) trained on CS-B		
Acceptance rate for human, true claim	96.8%	97.2%
Acceptance rate for synth, matched claim	2.5%	2.5%

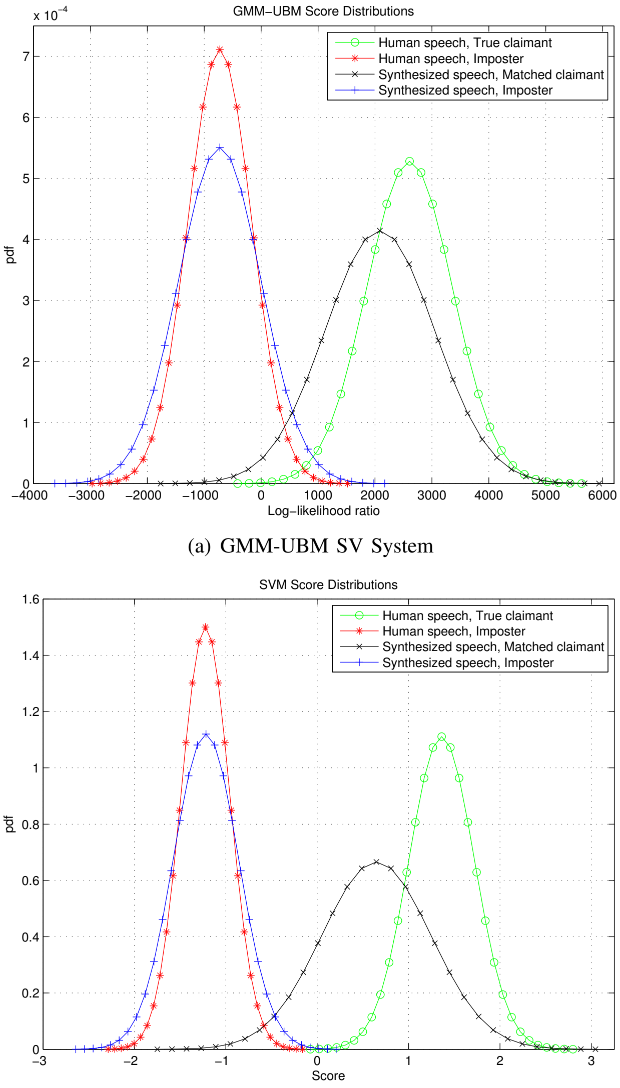
TABLE III
ACCURACY RATES FOR CLASSIFICATION OF HUMAN AND SYNTHETIC SPEECH. CLASSIFIER IS TRAINED WITH HUMAN SPEECH HS-B AND EITHER TTS-B OR CS-B FOR SYNTHETIC SPEECH. CLASSIFIER IS TESTED USING HS-C AND TTS-C. RESULTS ARE BASED ON A ZERO THRESHOLD FOR LOG-LIKELIHOOD RATIO (12) AND INCLUDE AN ADDITIONAL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

Training Data	Human Speech (HS-C)	Synthetic Speech (TTS-C)
HS-B/TTS-B	100%	100%
HS-B/CS-B	100%	90.10%
HS-B/CS-B (EER)	97.17%	97.17%

C. Evaluation of Sensitivity of Synthetic Speech Detector

In the evaluation of the SSD in Section VI-B, we have assumed that the same vocoder (STRAIGHT) and phase model (minimum phase) have been used in both training and test stages. Although STRAIGHT is the most popular approach to vocoding and the minimum phase model is normally used, in a real scenario, a different type of vocoder (e.g. [**?**]) and phase model could be used for imposture. Therefore we have investigated sensitivity to vocoder mismatch by experimenting with a simple vocoder which uses pulse/white noise excitation and the MLSA filter [**?**], [**?**]. We have also investigated sensitivity to phase model mismatch by experimenting with group delay modification [**?**].

Because the SSD features are entirely phased-based, any mismatch between vocoder and phase model which produces



(a) GMM-UBM SV System

(b) SVM using GMM supervectors SV system

Fig. 6. Approximate score distributions for (a) GMM-UBM and (b) SVM using GMM supervectors SV systems with human and synthesized speech. Distributions for human speech, true claimant (green lines, o) and synthesized speech, matched claimant (black lines, x) have significant overlap leading to a 81% acceptance rate for synthetic speech with matched claims.

different phase characteristics, may render the classifier's ability to detect synthetic speech unreliable. We have observed this effect in informal tests. When we train the SSD using the aforementioned vocoders, the accuracy of synthetic speech detection falls from 90.1% obtained with the original STRAIGHT vocoder, to 6.3% when training with the pulse/white noise excitation vocoder, and to 50% when training with the group delay modification vocoder. In all cases, the tests were done using TTS-C. On the other hand, classifier accuracy for the human speech remains at 100%. In order to address this issue, future research of a vocoder-independent and phase-adaptive approach such as MAP adaptation of the RPS-GMMs used for the SSD system, will have to be undertaken.

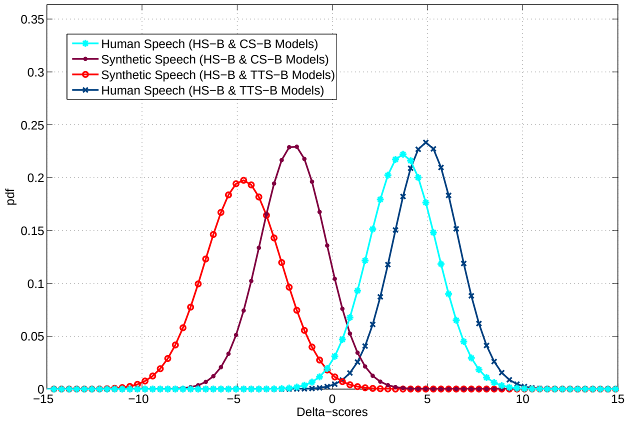


Fig. 7. Approximate distributions for the classifier scores, $\Lambda_{\text{RPS}}(\mathbf{Y})$ when tested with human and synthetic speech. The models for Human Speech are trained with HS-B. Blue and red curves show the classifier performance when the models for synthetic speech are trained using TTS-B. Cyan and magenta curves show the classifier performance when the models for synthetic speech are trained using coded speech CS-B. Both classifiers were tested with human speech HS-C and synthetic speech TTS-C.

D. Evaluation of Overall System

Next, we evaluated the *overall* system which includes the SV and SSD systems as illustrated in Fig. 5. Using the proposed SSD trained on TTS-B, we see in Table II rows 5-6, there is only a slight 0.1% drop to 99.6% in the acceptance rate for human speech for the GMM-UBM system and no change with the SVM system while the acceptance rate for synthetic speech is now reduced to 0% from over 81% thus clearly illustrating the effectiveness of the SSD using RPS features.

Training the SSD on CS-B, we see in Table II no change in the acceptance rate for human speech compared to training with TTS-B and an acceptance rate for synthetic speech of 8.8% for both SV systems. Finally, adjusting the decision threshold in the SSD for EER, we see in Table II a reduction in acceptance rate for synthetic speech to 2.5% with a slight decrease in acceptance rate for human speech (around 97%). 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, with only a slight loss in SV accuracy for human speech. Thus the proposed SSD using RPS features is an accurate and effective method for securing the SV systems against imposture using synthetic speech.

E. Evaluation of an Integrated System

Essentially Fig. 5 represents a system consisting of two separate classifiers: SV using MFCC features and SSD using RPS features. These classifiers can be integrated into a single classifier which uses vectors composed of both MFCC and RPS features. We extracted 53-D feature vectors by concatenating the MFCC feature vector (32-D) described in Section II-A with the RPS feature vector (21-D) described in Section IV-B. In the first simulation, the GMM-UBM and SVM classifiers based on the MFCC-RPS feature vectors were trained using HS-B only and in the second simulation, were trained using both HS-B and TTS-B datasets. When using HS-B and TTS-B, synthetic speakers were treated as imposters in the training stage. The systems were evaluated using HS-C and TTS-C datasets and the results are shown in Table **??**.

To begin the evaluation of the integrated system, we first must establish whether the addition of RPS features compromises SV accuracy when the system is trained and tested only with human speech. For this case, the addition of the RPS features slightly raises the EER to 0.35%, 0.02% for the GMM-UBM, SVM system, respectively as compared to the SV system which uses MFCC features only. The acceptance rates for true claims (99.7% for GMM-UBM, 100.0% for SVM) remain the same as compared to the SV system which uses MFCC features only. These results thus demonstrate that the addition of the RPS features does not appreciably change SV accuracy under human speech.

Earlier, we illustrated the imposture problem by demonstrating that when the SV systems using MFCC features were trained on human speech and tested with synthetic speech, the acceptance rates for matched claims were high (85.5%, 81.3% for the GMM-UBM, SVM systems, respectively). With the integrated system (GMM-UBM classifier), the acceptance rate for matched claims increases to 88.7% from 85.5%. On the other hand, the SVM system shows a notable drop in the acceptance rate to 56.2% from 81.3%. Unfortunately, both acceptance rates for synthetic speech with matched claims are still unacceptably high.

Next, we compare the integrated system trained with human and synthetic speech to the system composed of separate SV and SSD stages in Fig. 5. When the integrated system is tested with human speech, the acceptance rates for true claims drops slightly to 99.3% for the GMM-UBM system and remains the same 100% for the SVM system. When the GMM-UBM integrated system is evaluated with synthetic speech, the acceptance rate for matched claims is 40.6%. Not surprisingly, the GMM-UBM integrated system appears to have an average performance with synthetic speech between the stand-alone rates of the SV using MFCCs (85.5%) and the SSD using RPS (0.0%). When the SVM integrated system is evaluated with synthetic speech, the acceptance rate for matched claims is 3.5% which is still higher than for the system composed of separate SV and SSD stages which is also 0.0% (Table II, row 6). For both GMM-UBM and SVM integrated systems, inclusion of synthetic speech signals in training lowers the acceptance rates for synthetic speech, matched claims by around 50% (from 88.7% to 40.6% for GMM-UBM and from 56.2% to 3.5% for SVM). However, these results demonstrate that the proposed system composed of separate SV and SSD classifiers (Fig. 5) performs better than the integrated system. Nevertheless, the performance of the integrated SVM system is notable in that it does not use a separate synthetic impostor model for each speaker as the separate SSD does. Since training with CS-B leads to a less accurate model for synthetic speech than with TTS-B (see Table II, rows 6, 9, and 12) and results for the integrated system trained with TTS-B are worse than with the separate system, the integrated system is not trained with CS-B and evaluated.

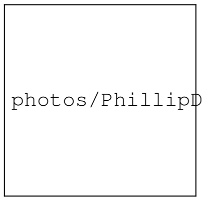
TABLE IV
ACCEPTANCE RATES FOR HUMAN SPEECH (TRUE CLAIMANT) AND SYNTHETIC SPEECH (MATCHED CLAIM) FOR THE INTEGRATED SYSTEM (SINGLE CLASSIFIER) WHICH USES VECTORS COMPOSED OF BOTH MFCC AND RPS FEATURES. IDEALLY THE SYSTEM HAS 100% ACCEPTANCE RATE FOR HUMAN SPEECH, TRUE CLAIM AND 0% FOR SYNTHETIC SPEECH, MATCHED CLAIM.

	GMM-UBM	SVM
Integrated System Trained on HS-B		
Acceptance rate for human, true claim	99.7%	100%
Acceptance rate for synthetic, matched claim	88.7%	56.2%
Integrated System Trained on HS-B and TTS-B		
Acceptance rate for human, true claim	99.3%	100%
Acceptance rate for synthetic, matched claim	40.6%	3.5%

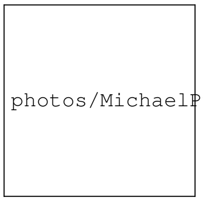
VII. Conclusions

In this paper, we have evaluated the vulnerability of speaker verification (SV) to imposture using synthetic speech. Using the Wall Street Journal (WSJ) corpus and two different SV systems (GMM-UBM and SVM using GMM supervectors), we have shown that with state-of-the-art speech synthesis, over 81% 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 voices to pass for true human claimants and hence poses a potential security problem.

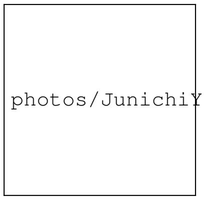
We have proposed a novel synthetic speech detector (SSD) based on relative phase shift (RPS) features. Although the SSD can detect human and synthetic speech with 100% accuracy, training requires that a TTS synthesizer be constructed for each speaker in the SV system which is not practical. Therefore, 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 synthetic speech, matched claims from over 81% to 2.5%, with a less than 3% drop in the acceptance rate for human speech, true claimants. However, the system is sensitive to the vocoder used: the same vocoder used by the impostor must be used to train the system. The investigation of vocoder independent techniques is left for future work.



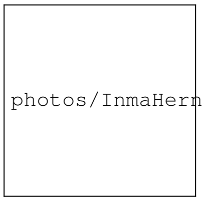
Phillip L. De Leon (SM'03) received the B.S. Electrical Engineering and the B.A. in Mathematics from the University of Texas at Austin, in 1989 and 1990 respectively and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Colorado at Boulder, in 1992 and 1995 respectively. In 2002, he was a visiting professor in the Department of Computer Science at University College Cork, Ireland. In 2008, he was selected by the U. S. State Department as a Fulbright Faculty Scholar and served as a visiting professor at Technical University in Vienna (TU-Wien). Currently, he is a Professor and Associate Department Head in the Klipsch School of Electrical and Computer Engineering and Director of the Advanced Speech and Audio Processing Laboratory at New Mexico State University. His research interests are in speech-signal processing, embedded systems, and pattern recognition and machine learning.



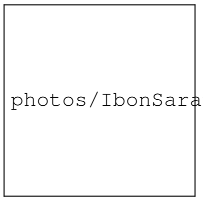
Michael Pucher is a senior researcher and project manager at the Telecommunications Research Center Vienna (FTW). He received a PhD from Graz University of Technology in 2007 with a thesis on semantic language modeling for speech recognition. His research interests are speech synthesis and recognition, multimodal dialog systems, and sensor fusion. He has authored and co-authored more than 30 refereed papers in international conferences and journals. In 2010 he was involved in the commercial development of Leopold, the first synthetic voice for Austrian German. In 2011 he was awarded a research grant from the Austrian Science Fund (FWF) for the project ”Adaptive Audio-Visual Dialect Speech Synthesis” (AVDS). A list of publications and a detailed CV can be found on http://userver.ftw.at/~pucher.



Junichi Yamagishi was awarded a Ph.D. by Tokyo Institute of Technology in 2006 with a thesis which pioneered the use of adaptation techniques in HMM-based speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of the Tokyo Institute of Technology in 2007. Since 2006, he has been in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. In addition to authoring and co-authoring over 80 refereed papers in international journals and conferences, his work has led directly to two large-scale EC FP7 projects and two collaborations based around clinical applications of this technology. In 2010, he was awarded the Itakura Prize (Innovative Young Researchers Prize) from the Acoustical Society of Japan for his achievements in adaptive speech synthesis. He is an external member of the Euan MacDonald Centre for MND Research in Edinburgh and a visiting associate professor of Nagoya Institute of Technology in Japan.



Inma Hernáez received the telecommunications engineering degree from the Universitat Politecnica de Catalunya (Spain), and the Ph.D. degree in telecommunications engineering from the University of the Basque Country (Spain), in 1987 and 1995, respectively. She is Full Professor in the Electronics and Telecommunication Department, Faculty of Engineering, University of the Basque Country, in the area of signal theory and communications and founding member and director of the Aholab Signal Processing Laboratory. Her research interests include signal processing and all aspects related to speech processing. She is also interested in the development of speech resources and technologies for the Basque language. She is a member of the International Speech Communication Association (ISCA), the Spanish thematic network on Speech Technologies (RTTH), and the European Center of Excellence on Speech Synthesis (http://www.ecess.eu).



Ibon Saratxaga received the telecommunications engineering degree from the University of the Basque Country (Spain), in 1995. From 2005 he has been a researcher at the Aholab Signal Processing Laboratory. He is currently teaching at the Faculty of Engineering in Bilbao. He has participated as junior research in several funded research projects. His research interests include speech coding and synthesis, with a focus on the study of the harmonic phase of the speech signal. He is a member of the International Speech Communication Association (ISCA), the Spanish thematic network on Speech Technologies (RTTH), and the European Center of Excellence of Speech Synthesis (http://www.ecess.eu).