DETECTION OF SYNTHETIC SPEECH FOR THE PROBLEM OF IMPOSTURE

Phillip L. De Leon
New Mexico State University
Klipsch School Elect. & Comp. Eng.
Las Cruces, New Mexico USA
pdeleon@nmsu.edu

Inma Hernaez, Ibon Saratxaga
Aholab SPL
Univ. of Basque Country
Bilbao, Spain
{inma.hernaez,ibon.saratxaga}@ehu.es

Michael Pucher
Telecommunications
Research Center
Vienna, Austria
pucher@ftw.at

Junichi Yamagishi
Centre for Speech Tech. Research
University of Edinburgh
Edinburgh, UK
jyamagis@inf.ed.ac.uk

ABSTRACT

In this paper, we present new results from our research into the vulnerability of a speaker verification (SV) system to synthetic speech. We use a HMM-based speech synthesizer, which creates synthetic speech for a targeted speaker through adaptation of a background model and both GMM-UBM and support vector machine (SVM) SV systems. Using 283 speakers from the Wall-Street Journal (WSJ) corpus, our SV systems have a 0.35% EER. When the systems are tested with synthetic speech generated from speaker models derived from the WSJ journal corpus, over 91% of the matched claims are accepted. We propose the use of relative phase shift (RPS) in order to detect synthetic speech and develop a GMM-based synthetic speech classifier (SSC). Using the SSC, we are able to correctly classify human speech in 95% of tests and synthetic speech in 88% of tests thus significantly reducing the vulnerability.

Index Terms— Speech synthesis, Speaker recognition, Security

1. INTRODUCTION

Synthetic speech potentially poses an imposture problem, that is a deception based on voice characteristics. An example is in remote or on-line authentication using speaker verification (SV), where a synthesized speech signal is substituted in order to wrongly gain access to person's account. In addition, synthetic speech also poses a potential problem when a SV system is used to confirm origination of a speech signal from a particular individual. In this case, the system might confirm origination when in fact the speech signal is synthetic. In both of these examples, the speech model for the synthesizer must be targeted to a specific person's voice.

Until recently, developing a text-to-speech (TTS) or speech 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 speech synthesis [1], the 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. In addition, recent experiments have also demonstrated that speaker-adaptive, HMM-based speech synthesis is robust to *non-ideal* training data. In [2] a high-quality voice was built from audio collected off of the Internet. The fact that small amounts of non-ideal training data can be used to construct high-quality synthetic speech, poses challenges to SV systems. The problem of imposture against SV systems using synthetic speech was first published over 10 years ago by Masuko, et. al. [3] and has been more recently studied in [4, 5] due to major advances in both SV and TTS systems.

This paper is a follow-on to our previous research on both the imposture problem and methods to detect synthetic speech. In [4], we utilized a very small corpus of HMM-based synthetic speech signals (9 speakers). Using a GMM-UBM SV system, we found that all synthetic speech signals were accepted as their human counterparts. We also proposed several methods to detect synthetic speech, including distance measures of dynamically time-warped MFCC features and error rates in automatic speech recognition systems. These approaches, however, were not able to consistently detect synthetic speech. In [5], we significantly expanded the corpus of synthetic speech signals to 283 speakers [based on the Wall-Street Journal (WSJ) corpus] and re-evaluated using the GMM-UBM system. We found over 90 of synthetic speech signals were accepted as their human counterparts. We also retested the proposed detection methods as well as a previously-proposed method, the average inter-frame difference of log-likelihood [6]. We found that all these methods failed to consistently detect synthetic speech using this corpus. This paper extends the work in the following ways. First, we have now implemented a state-of-the-art SV system based on support vector machine (SVM) using Gaussian supervectors [7] and evaluated it for synthetic speech signals. Second, we propose a promising method to detect synthetic speech based on relative phase shift (RPS) features.

This paper is organized as follows. In Section 2, we describe how we partition the WSJ corpus in order to train the TTS and SV systems and test the SV system for baseline (human speech) results. In Section 3, we briefly describe the SV systems used in our research. In Section 4, we describe the evaluation and provide results using the Wall-Street Journal (WSJ) corpus and its synthesized counterpart. In Section 5, we describe the RPS-based approach for detecting synthetic speech and provide preliminary results. Finally, we conclude the article in Section 6.

2. DATA SETS

As in [5], we use the WSJ corpus from LDC [8]. Although the WSJ corpus is not usually used for evaluating SV systems, it contains several hundred speakers and sufficiently long signals required for training both the TTS and SV system [8]. From the corpus, we chose the pre-defined official training data set (known as SI-284) that includes both WSJ0 and WSJ1. The SI-284 set has a total of 81 hours of speech data uttered by 284 speakers, however, one speaker was eliminated due to a poor recording resulting in 283 speakers. The material was partitioned into three sets A, B, and C. Referring to the second row in Table 1, Set A was used to train the TTS system, i.e., constructing an average voice model and for adapting the average voice model to the 283 target speakers. The details of the TTS systems used are described in [2, 5] Since the recording durations are variable in the WSJ corpus, we trained the TTS system using different durations. Set B was used to train the SV system, i.e. constructing the UBM and MAP-adapting the UBM as well as train the synthetic speech classifier (SSC). Note that the average voice model and UBM are trained on different subsets from the same corpus, since we aim avoid cross-corpus negative effects. However, in practice both the average voice model and UBM should be derived from different corpora; due to our limited access to appropriate speech corpora we were not able to achieve this. Set C was used for testing the SV and testing the SSC under human speech. Referring to the third row in Table 1, Set B was used to generate synthetic speech for training the SSC and Set C was used to generate synthetic speech for testing the SV system and the SSC system.

Table 1. Partitioning of the Wall Street Journal (WSJ) corpus into datasets used in the research to train the text-to-speech (TTS) system, train/test the speaker verification (SV) system, and train/test the synthetic speech detection (SSC) system.

	Set A	Set B	Set C
Human speech (from dataset)	train TTS	train SV train SSC	test SV test SSC
Synthetic speech (generated from dataset)		train SSC	test SV test SSC

3. SPEAKER VERIFICATION SYSTEMS

Our SV systems are based on the well-known GMM-UBM described in [9] and the support vector machine using Gaussian supervectors described in [7]. We briefly review these systems and our implementation.

3.1. 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 SV system is composed of two stages, shown in Fig. 1(a) and (b). The SVM using Gaussian supervectors SV system also 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 $1 \leq i \leq M$. 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 MAP-adapt only the mean vectors of the GMM-UBM to form speaker- and utterance-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, the mean vectors $\mu_{s,u,i}$ are then diagonally-scaled according to

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

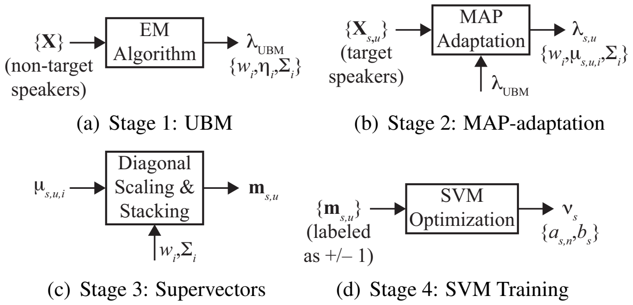

Fig. 1. Stages of training the SV systems. The GMM-UBM SV system is trained with (a)-(b) and the SVM SV system is trained with (a)-(d). Although the GMM-UBM is normally derived from non-target speakers, as described in Section 2, we have used target speakers.

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

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

In the fourth stage, 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 [7], 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.

3.2. System Testing

In 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

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

and accept the claim if

$$\Lambda(\mathbf{X}) \geq \theta \qquad (4)$$

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

$$y(\mathbf{X}) \quad = \quad \sum_{n\in\mathcal{S}} a_{C,n} t_{C,n} \mathbf{m}_{\text{test}}^T \mathbf{m}_{C,n} + b_C \qquad (5)$$

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

Table 2. SV system results.

	GMM-UBM	SVM
EER (human speech)	0.35%	0.35%
min DCF (human speech)	4.04e-3	2.36e-3
accepted claims from synthetic speech	259/283 = 91.5%	271/283 = 95.8%

3.3. Performance

The above systems have been implemented and tested using the NIST2002 one-speaker detection using cellular data. Following the evaluation protocol, the GMM-UBM system has 11.3% EER and 0.112 min DCF while the SVM system has 11.7% EER and 0.113 min DCF.

4. EXPERIMENTS AND RESULTS

Our simulations and tests are based on the WSJ corpus (described in Section 2 as well as the synthetic speech signals for target speakers generated from this corpus. For the GMM-UBM system we have trained on ≈90s speech signals and tested using ≈30s signals. Training signals for the SVM system were segmented into eight utterances per speaker. Test results for human speech signals are given in rows 2 and 3 of Table 2. The simulations for human speech were designed so that each test utterance has an associated true claim and 282 false claims yielding 283^2 tests. The unrealistically low EERs (0.35%) are due to the ideal nature of the recordings in the WSJ corpus and the accuracy of the SV systems. We note that both systems have about the same performance.

The simulations for synthetic speech were designed so that each test utterance has an associated matched claim yielding 283 tests. In a realistic imposture scenario, an attacker will generate a synthetic speech signal targeted at a specific speaker and make a claim only for that speaker, i.e. matched claim. Using the decision process outlined in Section 3.2, the systems were tested using the synthetic test signals and a matched claim of identity resulting in 283 tests. For the GMM-UBM system, the decision threshold is chosen as that for EER under human speech signal tests. Row 4 of Table 2 shows the results where we see over 90% of synthetic speech signals with an associated matched claim, will be accepted by the systems. Of particular interest in this paper, is the result that the SVM using Gaussian supervectors accepts even more claims using synthetic speech than the GMM-UBM despite both systems having the same performance using human speech. As we described in our earlier papers and verified with this new work, significant overlap occurs in the score distributions for human speech and synthetic speech. Thus adjustments in decision thresholding or standard score normalization techniques cannot differentiate between true and matched claims originating from human and synthesized speech.

5. DETECTION OF SYNTHETIC SPEECH USING RELATIVE PHASE SHIFT

Acoustic differences between the human and synthetic speech signals are audible even though the synthetic speech is of high-quality. Informal listening tests suggest human listeners can easily and consistently detect synthetic speech. Our previous attempts at automatic detection of synthetic speech have not been successful [4, 5]. In this work, we propose a new method based on differences in the relative phase shift (RPS) [10] between human and synthetic speech signals. We summarize RPS below.

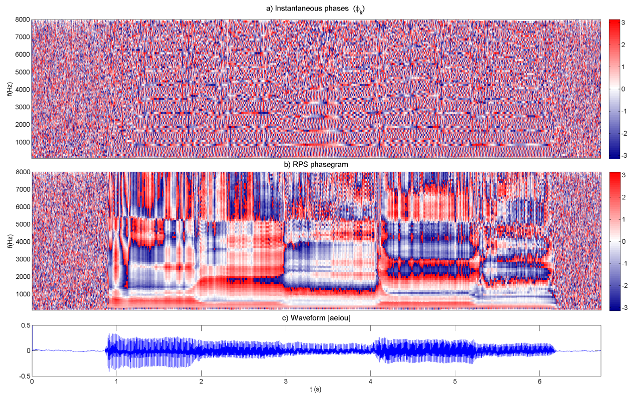

Fig. 2. Phasegrams of a voiced speech segment for five continuous vowels. a) Instantaneous phases, b) relative phase shift, and c) signal waveform.

5.1. Relative Phase Shift

RPS turns instantaneous phases from two harmonics into a relative measure against a common reference. This relative measure removes the linear phase component and allows phase structure to be more easily elucidated. To derive the RPS, we assume a harmonic model for voiced segments of the form

$$s(t) = \sum_k A_k(t)\cos[\varphi_k(t)] \tag{6}$$

where A_k are the amplitudes and

$$\varphi_k(t) = 2\pi k f_1 t + \theta_k \tag{7}$$

is the instantaneous phase of the kth harmonic; we denote the fundamental frequency as f_1 and the initial phase as θ_k. The instantaneous phase depends on the time instant and harmonic (encapsulated in the linear phase term $2\pi k f_1 t$). The initial phase shift θ_k is constant regardless of the time instant while the (periodic) waveform shape is stable (local stationarity assumption). Thus the waveform shape depends only on the differences between the initial phase shifts.

If we assume the fundamental as our reference and for simplicity $\theta_1 = 0$, we can solve for θ_k (now called the RPS) by equating the analysis time instants t_a in (7) between the kth harmonic and the reference fundamental:

$$\frac{\varphi_1(t_a)}{2\pi f_1} = \frac{\varphi_k(t_a) - \theta_k}{2\pi f_k} \tag{8}$$

or

$$\theta_k = \varphi_k(t_a) - k\varphi_1(t_a). \tag{9}$$

For a voiced segment, Figure 2(a) shows the instantaneous phase and Figure 2(b) shows the RPS where in the latter, we more easily see the phase-related structure.

In order to properly parameterize the RPS values to discriminate between human and synthetic speech, three important issues were addressed [11]:

- Due to the variable number of harmonics found in a predefined frequency range, the RPS vector has a varying dimension. This problem is solved by mel-filtering the RPS values with a constant number of filters.
- In order to avoid discontinuities, RPS values are unwrapped. The resulting envelope is differentiated in order to avoid ambiguity problems due to unwrapping.
- The Discrete Cosine Transform (DCT) is also used to further reduce the dimensionality.

5.2. RPS Modelling and Detection of Synthetic Speech

We modeled RPS of human and synthetic speech using speaker-dependent GMMs. In particular, 20 RPS-based features (differentiated and unwrapped RPS) have been computed every 10 ms over a 4 kHz bandwidth for voiced speech segments. The vector mean is removed before the DCT and appended as a element, resulting in 21 coefficients per RPS feature vector. We used 32 component densities in the GMMs and trained the classifier with varying lengths of voiced signal segments using Set B in Table 1. The human/synthetic speech decision is simply based on the log-likelihood ratio

$$\Lambda_{\text{RPS}}(\mathbf{X}_s) \quad = \quad \log p(\mathbf{X}_s|\lambda_{\text{human}_s}) - \log p(\mathbf{X}_s|\lambda_{\text{synth}_s}) \quad (10)$$

where $\mathbf{X}_s$ represents the sequence of RPS feature vectors for speaker s, λ_{human_s} is the GMM for human speech RPS features for speaker s, and λ_{synth_s} is the GMM for synthetic speech RPS features for speaker s. The classification is performed after speaker verification, where a hypothesis for speaker identity is available and the approriate synthetic and human speech models are selected based on this hypothesis. The speech signal is classified as human if $\Lambda_{\text{RPS}}(X_s) > 0$, otherwise it is classified as synthetic.

In our research, we tested the classifier using the human and synthetic speech speech signals from Set C in Table 1 after the SV system accepts a claim. The accuracy to detect human speech using the SSC trained using 10 s of voiced segments per speaker is 95% while that for synthetic speech is 88%. This is a marked improvement over our previous attempts to detect synthetic speech. On the other hand, the SSC incorrectly classifies about 4.2% of human speech as synthetic.

6. CONCLUSIONS

In this paper, we have evaluated two speaker verification systems, 1) Gaussian Mixture Model-Universal Background Model (GMM-UBM) and 2) support vector machine (SVM) using Gaussian supervectors using 283 human speech signals from the Wall Street Journal corpus and 283 synthetic speech signals derived from the corpus. While both systems have very low equal error rates (EERs), when presented synthetic speech, the GMM-UBM system accepts 92% of matched claims and the SVM accepts 96% of matched claims. Thus the speaker similarity/identity of synthetic speech is high enough to allow these synthesized voices to pass for true human claimants. These results suggest that high-quality synthetic speech may pose security issues for speech-based remote/online authentication or incorrect identity confirmation from a speech signal. We proposed a GMM-based classifier to detect synthetic speech based on the relative phase shift of voiced speech segments. Our results show we can detect synthetic speech up to 88% of the time although at the same time, 4.2% of the human speech will be incorrectly classified as synthetic.

Acknowledgements

MP is employed by the Telecommunications Research Center Vienna (FTW) which is supported by the Austrian Government and the City of Vienna within the competence center program COMET. JY is partially supported by EPSRC and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project). This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk). IH and IS have been partially supported by the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Berbatek, IE09-262).

7. REFERENCES

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

[2] 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.

[3] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. EUROSPEECH*, 1999.

[4] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, USA, 2010.

[5] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, Brno, Czech Republic, 2010.

[6] 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.

[7] 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.

[8] "Wall Street Journal Corpus," 2010.

[9] 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.

[10] Ibon Saratxaga, Inma Hernaez, Daniel Erro, Eva Navas, and Jon Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, pp. 381–383, 2009.

[11] Ibon Saratxaga, Inma Hernaez, Igor Odriozola, Eva Navas, Iker Luengo, and Daniel Erro, "Using harmonic phase information to improve ASR rate," in *Interspeech*, Japan, 2010.