REVISITING THE SECURITY OF SPEAKER VERIFICATION SYSTEMS AGAINST IMPOSTURE USING SYNTHETIC SPEECH

Phillip L. De Leon, Vijendra Raj Apsingekar — New Mexico State University, Klipsch School of Elect. & Comp. Eng., Las Cruces, New Mexico USA, {pdeleon,vijendra}@nmsu.edu

Michael Pucher — Telecommunications Research Center (FTW), Vienna, Austria, pucher@ftw.at

Junichi Yamagishi — Centre for Speech Technology Research (CSTR), Edinburgh, UK, jyamagis@inf.ed.ac.uk

ABSTRACT

In this paper, we investigate imposture using synthetic speech. Although this problem was first examined over a decade ago, dramatic improvements in both speaker verification (SV) and speech synthesis have renewed interest in this problem. We use a HMM-based speech synthesizer which creates synthetic speech for a targeted speaker through adaptation of a background model. We use two SV systems: standard GMM-UBM-based and a newer SVM-based. Our results show when the systems are tested with human speech, there are zero false acceptances and zero false rejections. However, when the systems are tested with synthesized speech, all claims for the targeted speaker are accepted while all other claims are rejected. We propose a two-step process for detection of synthesized speech in order to prevent this imposture. Overall, while SV systems have impressive accuracy, even with the proposed detector, high-quality synthetic speech will lead to an unacceptably high false acceptance rate.

Index Terms— Speech synthesis, Speaker recognition, Security

1. INTRODUCTION

The objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample [1]. During the training stage speaker-dependent feature vectors, based on mel-frequency cepstral coefficients (MFCCs), are extracted from training speech signals. Feature vectors from all users are concatenated and modeled with a Gaussian mixture model-universal background model (GMM-UBM), $\lambda_{\rm UBM}$ [1]. Next, individual speaker models, λ_s are constructed through MAP-adaptation of the GMM-UBM [1]. Both $\lambda_{\rm UBM}$ and λ_s are each parameterized by the set $\{w_i, \mu_i, \Sigma_i\}$ where w_i are the weights, μ_i are the mean vectors, and Σ_i are the diagonal covariance matrices of the GMM. During the testing stage feature vectors $\mathbf{x}_m$ are extracted from a test signal and a log-likelihood ratio $\Lambda(\mathbf{X})$ is computed by scoring the sequence of test feature vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ against the claimant model, λ_C and $\lambda_{\rm UBM}$

$$\Lambda(\mathbf{X}) \quad = \quad \log p(\mathbf{X}|\lambda_C) - \log p(\mathbf{X}|\lambda_{\rm UBM}). \qquad (1)$$

The claimant speaker is accepted if

$$\Lambda(\mathbf{X}) \geq \theta \qquad (2)$$

or else rejected, where θ is the decision threshold.

Support Vector Machines (SVMs) are binary discriminative classifiers that have more recently been applied to SV. The use of kernel functions in speech-related applications has led to many sequence kernels [2] and in particular, the GMM-supervector kernel has been successfully used in SV. The GMM-supervector kernel is defined as

$$K(\lambda_X, \lambda_C) \quad = \quad \sum_{i=1}^{W} w_i \mu_i^X \Sigma_i^{-1} \mu_i^C \qquad (3)$$

where W is the number of component densities in $\lambda_{\rm UBM}$, λ_X is the MAP-adapted (mean vectors only) model from test feature vectors $\mathbf{X}$, μ_i^X and μ_i^C are the respective mean vectors of λ_X and λ_C, and w_i and Σ_i are weights and covariance matrices of $\lambda_{\rm UBM}$. The SVM classifier is then based on (3).

Synthetic speech potentially poses two related problems for SV systems. The first problem is confirmation of an acquired speech signal as having originated from a claimed individual. In this case, a synthesized speech signal might be confirmed as having originated from an individual when it has not. The second problem is in remote or on-line authentication. In this case, a synthesized speech signal could be used to wrongly gain access to person's account. We assume for this second problem a text-prompted SV which does not present a problem for a speech synthesizer. In both of these problems, the speech model for the synthesizer must be targeted or matched to a specific person's voice.

The problem of imposture against SV systems using speech synthesized from hidden Markov models (HMMs) was first published over 10 years ago by Masuko, et. al. [3]. In their original work, the authors used an HMM-based text-prompted SV system [4] and an HMM-based speech synthesizer. In the SV system, feature vectors were scored against

speaker and background models composed of concatenated phoneme models (not GMM-based models). The authors also used a HMM-based speech synthesizer which was adapted to each of the human speakers [5].

When tested with 20 human speakers, the system had a 0% False Acceptance Rate (FAR) and 7.2% False Rejection Rate (FRR) and when tested with synthesized speech (20 synthetic voices) the system had over 70% FAR. In subsequent work by Masuko, et. al. [6], 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. By improving the SV system, the authors were able to lower the FAR for synthetic speech to 32%, however, the FAR for the human speech increased to 1.8%.

In the last 10 years, both SV systems and speech synthesizers have improved dramatically. Around the time as Masuko's work, GMM-UBM-based SV was proposed [1] which has been the standard method due to low equal-error rates (EERs). Other kernel-based techniques have been recently proposed and in some cases can lead to lower EERs [2]. HMM-based speech synthesizers have also improved in many ways yielding more natural-sounding speech. In addition, speaker models 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. Taken together, state-of-the-art speech synthesizers pose major challenges to state-of-the-art SV systems.

This paper is organized as follows. In Section 2, we describe our speech synthesis system. In Section 3, we describe our speaker verification systems. In Section 4, we describe the experimental evaluation and provide results using the human and synthesized speech corpus. Finally, we describe a concept for the detection of synthesized speech in Section 5 and conclude the article in Section 6.

2. SPEECH SYNTHESIZER

All adapted synthesizers described here are built using the framework from the speaker-adaptive "HTS-2007/2008" system [7]. The models were adapted from a background speaker model, which was trained with the speaker-dependent training methods described in [8]. The whole HMM-based speech synthesizer, illustrated in Fig. 1, comprises three main components: speaker-dependent training, speaker adaptation, and speech synthesis. In the speaker adaptation part, the speaker-dependent multi-stream left-to-right multi-space distribution hidden semi-Markov models (MSD-HSMMs) are transformed using constrained structural maximum a posteriori linear regression. In the speech generation part, 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 [9].

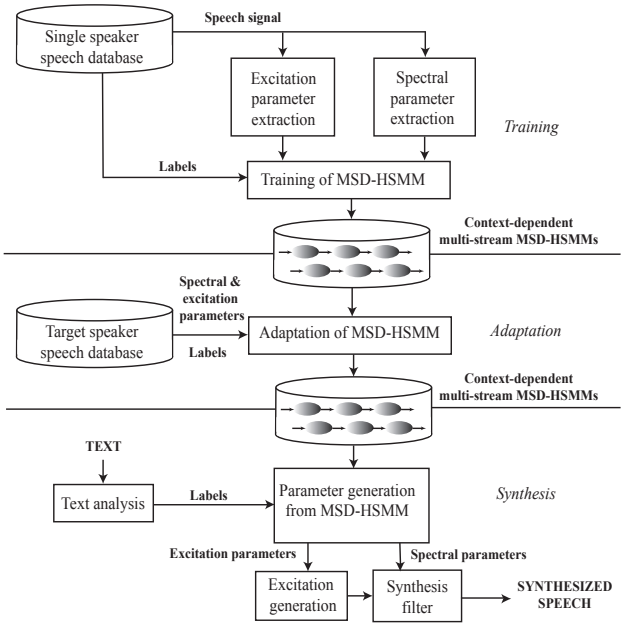

Fig. 1. HMM-based speech synthesis system which is a mixture of a speaker-dependent and speaker-adaptive systems.

3. SPEAKER VERIFICATION SYSTEM

Our GMM-UBM SV system uses feature vectors extracted every 10ms using a 25ms hamming window. The vector elements are 15 MFCCs, 15 Δ-MFCCs, and log- and Δ-log energy. We apply feature warping to the feature vectors in order to improve robustness. We use the Expectation Maximization (EM) algorithm to compute the parameters of the GMM-UBM and individual speaker models are obtained through MAP-adaptation of the GMM-UBM (only the mean vectors). Our 1024 component density, GMM-UBM SV system has baseline results as follows. For the 630 speaker TIMIT corpus (clean speech), we record 0.11% equal-error rate (EER) and for the 330 speaker NIST 2002 corpus (one speaker detection cellular task), we record 11.95% EER. Our SVM GMM-supervector SV system is based on the same parameters as the GMM-UBM SV system. The baseline EER is 8.0% for NIST 2002 corpus (100 speakers' test signals). These EERs closely agree with published values [10], [2].

4. EXPERIMENTS AND RESULTS

We recorded speech material for 10 human subjects in near-ideal recording conditions. The subjects were Austrian citizens and speak in the German language. This material was partitioned into three sets A, B, and C: set A was used for training the speech synthesizer background model, set B was used for training adapted synthetic voices, and set C was used for the SV system. For the synthesizer, we used training material (set A) from one speaker to create the background model

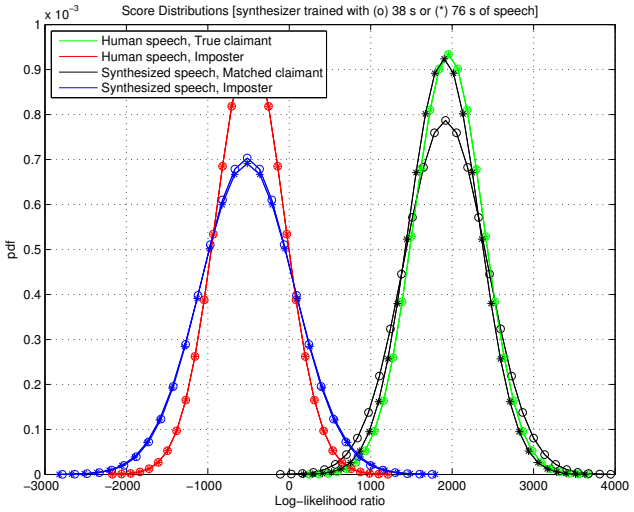

Fig. 2. Approximate score distributions for GMM-UBM SV system with human and synthesized speech. Distributions for synthesized speech (black and blue lines) nearly match those for human speech (green and red lines) leading to successful imposture using synthesized speech.

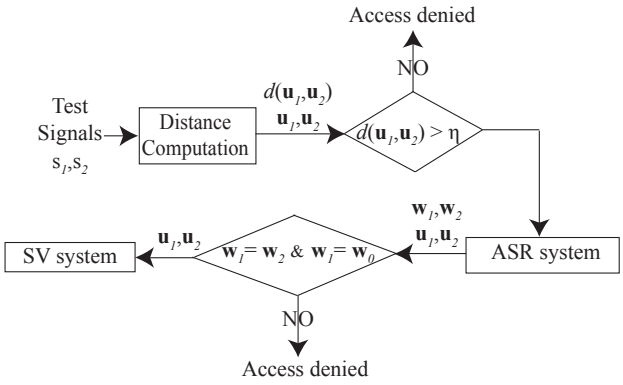

Fig. 3. Proposed system for detection of synthesized speech prior to speaker verification. System is composed of both MFCC distance measures of repeated utterances and ASR.

and adapted this model using the other 9 speakers' material from set B. We chose this procedure using one background speaker instead of training an average voice from multiple speakers to make an optimal use of our data set. We used varying lengths (19, 38, and 76 s) of training signals from set B to adapt the models. The adapted models were used to create synthesized speech for each of the 9 speakers (test signals for the SV). Due to the synthesizer complexity, we are limited in the number of speakers we can support in this research.

The one speaker used in building the background model for the synthesizer is intentionally left out of SV experiments so that the SV system would not be composed of any speakers in the background model. This is an important step since in a real-world SV system, the UBM and background model are unlikely to contain any common speakers. For the SV system, set C data from each of 9 speakers is split into 44 s training and 11 s test signals. A 256 component density GMM-UBM is computed from the 9 speakers' training data and individual speaker models are then adapted from the GMM-UBM. These same models are used in the SVM GMM-supervector system. The SV system is tested using the 9 human test signals and 27 synthesized speech signals (generated from models based on 19, 38, and 76 s of training data). Each of the 36 test signals is scored against one of 9 (human) claimants leading to a total of 324 tests.

When the SV systems (GMM-UBM and SVM GMM-supervector) are tested with human speech there are zero false acceptances (FAs) and zero false rejections (FRs), thus the systems each perform perfectly. This is to be expected given the performance of our systems with ideal/near-ideal recordings. The mean and variance of the log-likelihood scores for the GMM-UBM SV system are computed and approximate score distributions for human speech are shown in Fig. 2 with green and red lines.

Next the SV systems are tested using synthesized speech. We find that when the synthesized voices claim to be their human counterparts, i.e. matched claimant, the system accepts the claim each time but rejects all other claims. Thus despite the state-of-the-art performance of the SV systems, the quality of the synthesized speech is high enough to allow these synthesized voices to pass for true human claimants. This is true even for the synthesizer trained with as little as 19 s of data. The mean and variance of the log-likelihood scores (1) are computed and approximate score distributions for synthesized speech (black and blue lines) are shown in Fig. 2. Most worrisome in this experiment is the mean and variance of the score distributions for true and matched claimants are nearly equal. Thus adjustments in decision thresholding or standard score normalization techniques are unlikely to differentiate between true and matched claims originating from human and synthesized speech.

5. DETECTION OF SYNTHESIZED SPEECH

As a consequence of our results, we propose a two-step system for the *detection* of synthesized speech, which is depicted in Fig. 3. The system relies on the high-degree of regularity of repeated utterances from a speech synthesizer and the higher error rates of an automatic speech recognizer (ASR) (trained on human speech) subjected to synthesized speech. The system parameters are the distance threshold η, reference utterance word string $\mathbf{w}_0$, and distance function $d(\{\mathbf{x}_1\}, \{\mathbf{x}_2\})$.

In the first step, we compute the acoustic distance between two realizations of the same utterance using dynamic time warping (DTW) of MFCC features. This exploits the fact that the HMM-based synthesizer will always produce the same

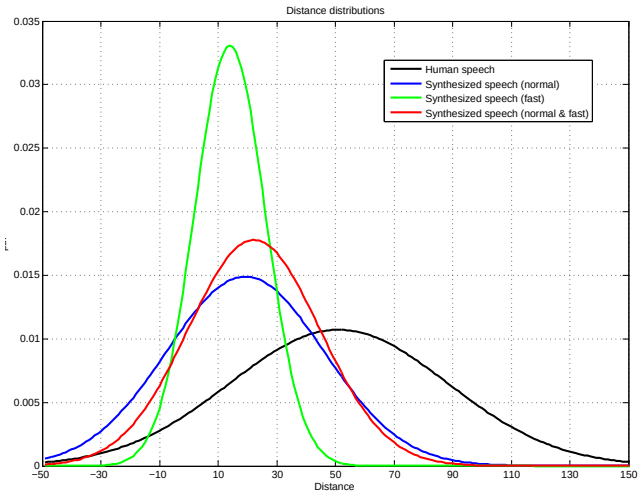

Fig. 4. Distributions of DTW distance of MFCCs for human and synthetic speech with different linguistic contexts and durations.

Table 1. Speech recognition WERs and SERs in %.

Dataset	Grammar1 (91 sentences)	Grammar2 (199 sentences)
Human speech	9.54 / 8.76	13.44 / 13.38
Synthetic (76 sec.)	11.64 / 10.82	15.62 / 16.09
Synthetic (38 sec.)	14.44 / 13.98	18.36 / 19.00
Synthetic (19 sec.)	26.50 / 29.52	31.33 / 36.16

globally optimal waveform in terms of maximum likelihood, given a set of input phoneme labels while human speech will always be different. In Fig. 4 we see that synthetic speech phrases are more similar to each other than human speech, even when using different linguistic contexts and durations, which means that small changes in the synthesis parameters are not sufficient to make synthetic speech less similar.

In the second step, we perform automatic speech recognition (ASR) on input utterances. This can prevent some FAs from synthesizers trained with small amounts of speech as shown by the word-error-rates (WERs) and sentence-eror-rates (SERs) in Table 1. During this step, we verify the same utterance was spoken twice ($\mathbf{w}_1 = \mathbf{w}_2$) and reference utterance was spoken ($\mathbf{w}_1 = \mathbf{w}_0$). With these steps, the SV system is able to prevent some impostures using synthesized speech.

6. CONCLUSIONS

In this paper, we have revisited the problem of imposture against speaker verification (SV) systems using synthetic speech. We used a HMM-based speech synthesizer where the model for a targetted speaker is adapted from a background model. We tested two different SV systems: a GMM-UBM system and SVM GMM-supervector system. Our results show that when the synthesized voices claim to be their human counterparts, i.e. matched claimant, the SV systems accept the claim each time. Next, we proposed a system for detection of synthetic speech based on acoustic distances between repeated utterances and speech recognition error rates. Despite the state-of-the-art performance of all systems involved, the quality of the synthesized speech is high enough to allow these synthesized voices to pass for true human claimants. This result suggests that high-quality synthetic speech may lead to a high false acceptance rate and may pose security issues for speech-based remote/online authentication or incorrect identity confirmation from a speech signal. We will evaluate this result in future work using broadcast news speech corpora with hundreds of speakers.

7. 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] C. Longworth and M.L.F. 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.

[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] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.