Preventing converted speech spoofing attacks in speaker verification

M. J. Correia[*,**], A. Abad[*,**] and I. Trancoso[*,**]
[*] Instituto Superior Técnico, Lisboa, Portugal
[**] INESC-ID/Spoken Language Laboratory, Lisboa, Portugal
{joana.correia,alberto.abad,isabel.trancoso}@l2f.inesc-id.pt

Abstract – Voice conversion (VC) techniques, which modify a speaker's voice to sound like another's, present a threat to automatic speaker verification (SV) systems. In this paper, we evaluate the vulnerability of a state-of-the-art SV system against a converted speech spoofing attack. To overcome the spoofing attack, we implement state-of-the-art converted speech detectors based on short- and long-term features. We propose a new converted speech detector using a compact feature representation and a discriminative modeling approach. We experiment pairing converted speech detectors based on short- and long-term features to improve converted speech detection. The results indicate that the proposed converted speech detector pair outperforms state-of-the-art ones, achieving a detection accuracy of 97.9% for natural utterances and 98.0% for converted utterances. We include the anti-spoofing mechanism in our SV system as a post-processing module for accepted trials and reevaluate its performance, comparing it with the performance of an ideal system. Our results show that the SV system's performance returns to acceptable values, with less than 1.6% equal error rate (EER) change.

I. INTRODUCTION

Speaker verification (SV) is the binary task of accepting or rejecting a claimed identity, based on a user's utterance [1]. This task falls under the broader category of biometrics and, as such, has many applications in access control systems, telephone banking, voice mail or calling cards [1][2]. The security of SV systems can be threatened by other speech processing techniques, particularly voice conversion (VC)[3], whose task is to modify the one speaker's voice characteristics, the source speaker, into sounding as if they were of another speaker, the target speaker, without changing the linguistic contents of the converted utterance. A converted utterance of the source speaker into the target speaker could then be used to try to attack or fool a security system.

The vulnerability of SV systems against spoofing attacks has been widely recognized [4][5]. In order to protect a SV system against a spoofing attack of synthetic nature (either converted speech or synthetic speech) and make it more robust, one may implement a synthetic speech detection module that discriminates between natural and synthetic speech. The development of synthetic speech detectors is a relatively recent research topic [6][7][8][9][10].

The goal of this study is to investigate the performance of SV systems against converted speech attacks. We start by evaluating the vulnerability of a state-of-the-art i-vector SV system against a GMM-based and a unit selection (US) -based converted speech corpus with telephone quality. Secondly we implement two state-of-the-art converted speech detectors as in [7][8] where the detectors are based on short- and long-term features, extracted from both the magnitude or the modified group delay function phase spectrum (MGDFPS). The features are used to train GMM-based detectors. Then we fuse the scores of the two systems using logistic regression optimization. Thirdly we propose a new converted speech detector trained with a compact representation of the features extracted previously and using a discriminative learning model, as we consider it more suitable for a binary discrimination task which is the case of converted speech detection. For this we adopted a support vector machine (SVM). The scores of the two converted speech detectors are fused as in the previous case. Finally we incorporate the best converted speech detectors in our SV system, as an anti-spoofing mechanism and reevaluate its performance, comparing it with the performance of the system without protection and to the performance of an ideal anti-spoofing mechanism.

This paper is organized as follows: section II describes the speaker verification system, based on total variability modeling, that was used to conduct this study; section III briefly describes the two voice conversion methods used to create our spoofing corpora; the state-of-the-art and the proposed converted speech detectors are presented in section IV; our experiments are described on section V and, lastly some conclusions are drawn on section VI.

II. SPEAKER VERIFICATION SYSTEM

In this study we consider a state-of-the-art SV system based on i-vectors, as proposed in [11].

A. I-vectors

I-vectos are based on total variability modeling, a technique that rapidly emerged as a powerful approach for SV and has become a current *de facto* standard. In this approach, closely related to the joint factor analysis (JFA), the speaker and channel variability of a high dimensional GMM supervector are jointly modeled as a single low-rank total-variability space. The low-dimensional total-variability factors are extracted from a speech segment to form a vector, called i-vector, which represents the speech segment in a compact and efficient way.

In our experiments, we computed the i-vectors based on MFCC features extracted in 20ms frames, updated every10ms. Each feature vector had 12 MFCC, log-energy and the corresponding velocities and accelerations,

totaling 39 dimensions. The total variability matrix was estimated according to [11] using a universal background model (UBM) composed by 1024 Gaussians that was trained using the 1conv4w-1conv4w training data subsets of NIST SRE 2004 and 2005 corpora. The dimension of the total variability sub-space was set to 400. No channel compensation or probabilistic linear discriminant analysis (PLDA) were applied, making this a simple SV system.

The verification score is obtained by simple cosine similarity between the target speaker i-vector and the test segment i-vector.

III. VOICE CONVERSION METHODS

In order to simulate the spoofing attacks to our SV system we considered two different voice conversion methods: GMM-based voice conversion, and US-based voice conversion.

A. GMM-based conversion

One of the most popular methods for voice conversion was originally proposed by [12] and is based on the joint density Gaussian mixture model.

This model requires *N*-dimensional time aligned acoustic features, $\mathbf{X} = [\mathbf{x}_1', \mathbf{x}_2', \dots, \mathbf{x}_N']'$, from the source speaker and $\mathbf{Y} = [\mathbf{y}_1', \mathbf{y}_2', \dots, \mathbf{y}_N']'$, from the target speaker, determined, for instance, by Dynamic Time Warping (DTW).

In the GMM algorithm, the joint probability function of the acoustic features is defined as:

$$\boldsymbol{P}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{P}(\boldsymbol{Z}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} \boldsymbol{\alpha}_{\boldsymbol{i}}^{(\boldsymbol{z})} \boldsymbol{\mathcal{N}}\left(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{i}}^{(\boldsymbol{z})}; \boldsymbol{\Sigma}_{\boldsymbol{i}}^{(\boldsymbol{z})}\right), \sum_{i=1}^{M} \alpha_i^{(z)} = 1, \alpha_i^{(z)} > 0 \quad (1)$$

where $\boldsymbol{\mu}_{\boldsymbol{i}}^{(\boldsymbol{z})}$ is the mean and $\boldsymbol{\Sigma}_{\boldsymbol{i}}^{(\boldsymbol{z})}$ is the covariance of the *M*-variate normal distributions. Parameters $\lambda^{(z)} = \left\{\alpha_i^{(z)}, \boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma}_i^{(z)} | i = 1,2,\dots,M\right\}$ are estimated using the estimation maximization (EM) algorithm [2].

The mapping function [13] used to convert features from the source speaker to target speaker is given by:

$$F(\boldsymbol{x}) = E(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{M} p_i(\boldsymbol{x}) \left(\boldsymbol{\mu}_i^{(z)} + \boldsymbol{\Sigma}_i^{(yx)} \left(\boldsymbol{\Sigma}_i^{(xy)}\right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_i^{(z)}\right)\right) \quad (2)$$

where $p_i(\boldsymbol{x}) = \frac{\alpha_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i^{\boldsymbol{x}}, \boldsymbol{\Sigma}_i^{\boldsymbol{xx}})}{\sum_{k=1}^{L} \alpha_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\Sigma}_k^{\boldsymbol{xx}})}$ is the posterior probability of the source vector belonging to the i^{th} mixture component.

B. Unit Selection-based conversion

Unlike GMM-based voice conversion, US-based voice conversion does not require parallel data between speakers; instead, it uses the target speaker's voice to directly synthesize new speech [14].

The goal of US is to, given a sequence of source speech features, $\boldsymbol{x}_1^M$, find the best fitting sequence of target speech features, $\boldsymbol{y}_1^M$, that minimizes the target cost (an estimate of the difference between the database unit u_m and the target t_m which it is supposed to represent), and the concatenation cost (an estimate of the quality of a join between the consecutive units u_{m-1} and u_m) [15]. The target vector sequence is given by:

$$y_1^M = \arg\min_{y_1^M} \sum_{m=1}^{M} \{\alpha S(\boldsymbol{y}_m - \boldsymbol{x}_m) + (1-\alpha) S(\boldsymbol{y}_{m-1} - \boldsymbol{y}_m)\}, \quad (3)$$

where α is a parameter to adjust the tradeoff between fitting the accuracy of source and target sequences and the spectral continuity criteria.

IV. CONVERTED SPEECH DETECTION MODULES

A good approach to build a converted speech detector is to train a model that characterizes the converted speech signal. The features used to train such a model should contain relevant information on the characteristics of converted speech vs. natural speech.

The most common methods for voice conversion use features derived from short-term magnitude spectrum to train and estimate the conversion. Hence, during the conversion process, information that is not from frame level and relative to the magnitude spectrum is left out. Particular examples of discarded information include information about the phase spectrum, and information about the evolution of the speech signal over time. As a consequence of this information loss, systematical artifacts are produced in the converted speech.

A. Information extraction

In [7][8] it is suggested that good features used to characterize converted speech may be extracted from the MGDFPS or from the temporal modulation of either the magnitude or the MGDFPS. The goal of using such features is to easily detect the artifacts on converted speech created by the information lost during the analysis-synthesis process of voice conversion.

1) Phase information

In order to extract features derived directly from the phase spectrum of a speech signal, it is necessary to compute the unwrapped phase [15]. An alternative that is computationally simpler is using the group delay function phase spectrum (GDFPS) [16], which has the additional advantage of reducing the effects of noise.

The GDFPS is a measure of non-linearity of the phase spectrum [17] and is defined as the negative derivative of the phase spectrum with respect to the frequency:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}, \quad (4)$$

Where $X(\omega)$ and $Y(\omega)$ are the short time Fourier transform (STFT) of $x(n)$ and $nx(n)$, $X_R(\omega)$, $X_I(\omega)$, $Y_R(\omega)$ and $Y_I(\omega)$ are the real and imaginary part of $X(\omega)$ and $Y(\omega)$, respectively.

Given a speech signal, the computation of group delay cepstral coefficients (GDCC) for each speech segment, $x(n)$, of 20ms, updated every 10ms was achieved as follows:

1. Computing the STFT $X(\omega)$ and $Y(\omega)$ of $x(n)$ and $nx(n)$, respectively.
2. Computing the GDFPS as in Equation (4).

3. Applying a 24 filter Mel-frequency filterbank to the MGDFPS to obtain filter-bank coefficients.
4. Applying the discrete cosine transform (DCT) to the filter-bank coefficients to obtain 12 GDCC.

The resulting 12 GDCC are used as feature vectors for model training. For comparison sake, tests will also include *de facto* standard features (12 MFCC, without deltas or delta-deltas, extracted from the same speech frames).

2) Temporal modulation information

To capture the correlation between frames and the temporal characteristics of features trajectories both in the magnitude spectrum and in the GDFPS, we compute the magnitude modulation (MM) features and the phase modulation (PM) features, respectively [8].

The MM were derived as follows:

1. Dividing the power spectrogram into 50 frame segments with a 30 frame overlap.
2. Applying a 20 filter Mel-filterbank to the spectrogram to obtain the filter-bank coefficients, forming an 20×50 matrix.
3. Applying mean variance normalization (MVN) to the trajectory of each filter-bank.
4. Computing the 64-point FFT of the 20 normalized trajectories.
5. Concatenating every modulation spectra to form a $20 \times 32 = 640$ coefficients modulation supervector.
6. Applying principal component analysis (PCA) to the modulation supervector to reduce dimensionnality and eliminate dimensions with high correlation. We kept the 10 projected dimensions with the largest associated variance.

The MM feature vectors are then 10-dimensional and are used as feature vectors for model training.

To derive the PM features, we followed the same steps, but applying them to the group delay function phase spectrogram instead of the power spectrogram.

B. Model training

In this study, we modeled the feature distributions with a generative model, the GMM, which is the preferred model for state-of-the-art converted speech detectors. We propose a new compact representation of the features that we model this using a discriminative approach, the SVM.

1) GMM model

Two GMM models are trained, one with natural data and another with converted data. The converted or natural decision is done based on the log likelihood radio:

$$L(O) = \log p(O|\lambda_{natural}) - \log p(O|\lambda_{converted}), \quad (5)$$

where O is the feature vector sequence of the test utterance, λ_{natural} and $\lambda_{\text{converted}}$ are the GMM models for converted and natural speech, respectively.

For the GMM models of the short-term features we studied (MFCC and GDCC), we adopted 512 Gaussian components to model the distributions. For the long-term features (MM and PM) we adopted 32 Gaussian components. We chose a smaller number of components for the latter features because those are extracted over a longer speech segment.

2) SVM

The converted or natural discrimination task is a binary task; as such, we considered studying the performance of a discriminative approach to address it. Over the last decade, SVM-based methods have been outperforming log likelihood ratio-based methods in SV problems [18], which further motivated us to try this approach.

Given a training set of labeled, two-class examples, an SVM estimates a hyperplane that maximizes the separation of the two classes, after transforming it to a high dimensional space via Kernel function. SVMs are constructed as a weighted sum of a kernel function:

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(x, v_i) + k, \quad (6)$$

where x is the input data, N is the number of support vectors, α_i and k are training parameters, v_i are the support vectors, obtained via an optimization process.

In this study we chose a linear kernel and sequential minimal optimization (SMO) as the optimization algorithm.

Traditionally, the output of an SVM for inputed test data is a predicted label, p.e. 0 or 1. This label is assigned depending on which side of the separating hyperplane do the test input features fall on. A more detailed output is the distance of test features to the hyperplane. We opted to use that distance as our SVM output in order to allow a finer score fusion for the two systems.

a) Compact feature representation

The previously described feature representation (used in the state-of-the-art converted speech detectors) resulted in a matrix of $N \times C$ coefficients for each speech file, where N is the number of frames in the file and C is the number of coefficients of the feature vectors. This meant that, for each minute of speech, with features extracted every 10 ms, we would have a $6,000 \times C$ matrix.

Alternatively to this consuming, full representation of information, we propose the use of a lighter feature representation. To use this feature representation, given a speech file, we do feature extraction as described and fit a normal curve to the distribution of each of the coefficients. We keep the fitted parameters, the average and standard deviation of each coefficient over the whole utterance and form a new feature vector, of dimension $1 \times 2C$, that compactly represents the speech file. Comparatively to the full representation we decrease the number of feature vectors approximately 10^4 times.

This representation conversion can be made after feature extraction or during the feature extraction process.

The greatest advantage of such representation is that it reduces the training time of the model from several hours to a few seconds.

C. Score Fusion

The features extracted from short-time features carry complementary information to the long term-features [7]. As such, it may be useful to fuse the scores of the two converted speech detectors in order to make the overall performance of the anti-spoofing mechanism more robust.

To perform the score fusion, we have used the fusion algorithms implemented in the BOSARIS toolkit for matlab [19], which performs logistic regression to fuse multiple sub-systems of binary classification.

V. Experimental Setup and Results

A. Corpora

In this study we use 4 main speech corpora, either taken directly from the NIST SRE2006 or derived from it: 1) 782 files of the training data in the core task, 1conv4w-1conv4w, of NIST SRE2006; 2) 3647 randomly chosen files of the test data in of the same subset of NIST SRE2006; 3) a set of 2904 files of GMM-based converted speech; 4) a set 2902 files of US-based converted speech.

The converted corpora used utterances from the NIST SRE2006 3conv4w and 8conv4w training sections as source data and the conversion matched randomly chosen same gender speakers from the 1conv4w-1conv4w of the NIST SRE2006.

Table I summarizes the usage of the available corpora for training and testing the several systems. We note that there is no file overlap on the training and testing sets.

TABLE I. CORPORA USED FOR MODEL TRAINING AND TESTING

		SRE2006 1conv4w-1conv4w train	SRE2006 1conv4w-1conv4w test	GMM-based converted speech	US-based converted speech
SV system	train	782			
	test		3647	2447	2449
converted detectors	train	300		300	300
	test		2459	2447	2449

B. Spoofing data against speaker verification system

As performance measure for the SV system we consider the EER.

The system was tested against 1458 natural genuine trials and 2700 natural impostor trials. To simulate spoofing attacks, we added to the test trials 2164 GMM-based converted impostor trials and 2196 US-based converted impostor trials. The EER is presented in Table II. The system detection error tradeoff (DET) curve is presented in Fig. 1.

Table II shows that the EER increased when the system was under a spoofing attack, a result in line with other previous studies. We note that our system is slightly more vulnerable to attacks with US-based converted data than with GMM-based converted data.

The increase of the EER is a result of the increased number of false acceptances (FA) of the system, a consequence of all the misclassifications of the converted impostor trials. The number of misses remained constant.

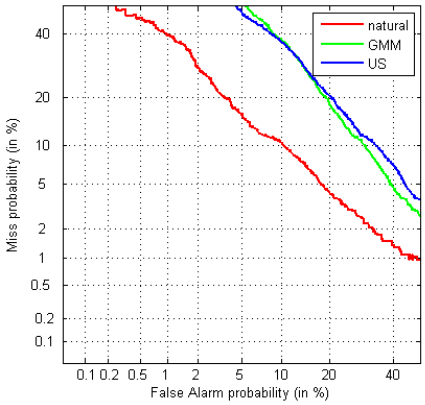

Fig. 1 DET curve of the SV system performance when against natural data; natural data and converted GMM impostors; natural data and US impostors

TABLE II. EER % OF SV SYSTEM PERFORMANCE AGAINST NATURAL DATA; NATURAL DATA AND GMM-BASED IMPOSTORS; NATURAL DATA AND US-BASED IMPOSTORS

Voice conversion method	SV system performance (EER %)
Baseline (no conversion)	9.4
GMM-based conversion	18.9
US-based conversion	20.0

C. Spoofing data against converted speech detectors

Firstly we evaluated the performance of the converted speech detectors separately, using as performance measure the system accuracy (Acc) percentage.

Every converted speech detector was trained and tested with the same train and test sets, for better comparability of the performance. The test set consisted of 2459 natural speech files, 2447 GMM-based converted speech files and 2449 US-based converted speech files. We measured the Acc% of the natural and converted files.

Table III summarizes the performance of the converted speech detector for each combination of model, short-term feature and training data. The results for the long-term features are presented in Table IV.

From Tables III and IV we can make the following observations: mixing converted train data (GMM-based and US-based) does not always improve the performance of the detector. Regarding the converted detectors trained with short-term features, we observed that the ones using GDCC over-performed the ones using MFCCs, a result in line with previous studies, that confirms the efficiency of discriminating natural and converted speech through searching for the artifacts created in the phase spectrum during the analysis-synthesis stage of the conversion.

It is unclear which of the long-term features performs better, given the similar accuracy for equivalent testing conditions. The systems averaged 68.7% for the PM modeled by a GMM, 68.7% for the MM modeled by GMM, 79.6% for the PM modeled by SVM and 80.1% for MM modeled by SVM, which confirms that the modulation features contain important information on the naturalness of the speech signal.

Comparing the performance of the two models, we observed that, with the exception of the model trained with MFCC as features, the proposed SVM out-performed the GMM models in the majority of the test scenarios. In particular, we highlight the SVM scores obtained by the detector trained with mixed converted data, using GDCC as features, which averaged 97.9% accuracy.

We proceeded to fuse the scores of our speech detectors that modeled short and long-term information. For this, we considered only relevant features, leaving out the MFCC as a consequence of poor performance. The fusions we considered were the following: 1) GDCC and MM for every condition; 2) GDCC and PM for every condition. The results relative to the score fusion are presented in Table V.

TABLE III. PERFORMANCE (ACC %)OF THE CONVERTED SPEECH DETECTORS USING SHORT-TERM FEATURES WITH VARIABLE TRAIN, TEST AND MODELING CONDITIONS

Short--term feature	Train data	Test data	GMM model		SVM	
			Natural trials (Acc%)	*Conv. trials (Acc%)*	*Natural trials (Acc%)*	*Conv. trials (Acc%)*
MFCC	GMM	GMM	87.0	89.5	75.2	54.9
		US	87.0	20.6	71.6	45.4
	US	GMM	85.4	95.3	68.0	74.9
		US	85.4	89.3	82.8	88.2
	mix	GMM	84.6	93.4	65.9	78.2
		US	84.6	79.8	70.6	79.0
GDCC	GMM	GMM	91.0	89.4	97.5	94.4
		US	91.0	52.1	97.9	80.0
	US	GMM	96.1	86.8	98.2	84.6
		US	96.1	74.0	82.1	93.9
	mix	GMM	91.5	92.3	97.7	98.0
		US	91.5	77.1	97.6	98.5

TABLE IV. PERFORMANCE (ACC %)OF THE CONVERTED SPEECH DETECTORS USING LONG-TERM FEATURES WITH VARIABLE TRAIN, TEST AND MODELING CONDITIONS

Long--term feature	Train data	Test data	GMM model		SVM	
			Natural trials (Acc%)	*Conv. trials (Acc%)*	*Natural trials (Acc%)*	*Conv. trials (Acc%)*
PM	GMM	GMM	73.1	78.8	75.1	83.4
		US	73.1	50.9	75.1	82.4
	US	GMM	73.1	78.8	82.9	61.5
		US	73.1	50.9	82.9	90.7
	mix	GMM	62.8	72.4	75.7	78.8
		US	62.8	74.5	75.7	91.1
MM	GMM	GMM	72.0	80.1	73.5	83.8
		US	72.0	50.7	73.5	84.1
	US	GMM	72.0	80.1	84.0	64.3
		US	72.0	50.7	84.0	90.6
	mix	GMM	62.2	75.1	76.7	78.6
		US	62.2	74.7	76.7	91.3

From Table V we observe that score fusion improved the accuracy percentage in most of the test conditions, yielding an average of 91.1% for the PM and GDCC fusion modeled by GMM, 91.0% for the MM and GDCC fusion modeled by GMM, 94.7% for the PM and GDCC fusion modeled by the proposed SVM, and 94.8% for the MM and GDCC fusion modeled by the proposed SVM.

The combination of feature pair, training data and model with the highest accuracy was the MM and GDCC, trained with mixed data and modeled by the SVM, scoring 98.0%. This was therefore the combination chosen to incorporate our converted detectors in the SV system.

TABLE V. PERFORMANCE (ACC %)OF THE FUSED CONVERTED SPEECH DETECTORS USING SHORT- AND LONG-TERM FEATURES WITH VARIABLE TRAIN, TEST AND MODELING CONDITIONS

Feature pair	Train data	Test data	GMM model		SVM	
			Natural trials (Acc%)	*Conv. trials (Acc%)*	*Natural trials (Acc%)*	*Conv. trials (Acc%)*
PM + GDCC	GMM	GMM	94.5	96.6	98.3	97.5
		US	78.0	79.8	92.4	91.8
	US	GMM	93.2	98.5	84.3	83.4
		US	84.0	94.5	98.0	98.6
	mix	GMM	96.3	95.7	98.1	97.2
		US	87.1	94.9	98.1	98.0
MM + GDCC	GMM	GMM	94.6	96.6	98.4	97.5
		US	77.3	79.2	92.1	91.6
	US	GMM	93.2	98.1	85.3	83.4
		US	84.4	94.9	98.7	98.6
	mix	GMM	96.6	96.0	98.2	97.8
		US	87.1	94.4	98.0	98.2

D. Spoofing data against speeaker verification system with anti-spoofing mechanisms

To reduce the effects of spoofing attacks in the performance of our SV system, particularly in the FA, we incorporated our best converted speech detector as an anti-spoofing mechanism as shown in Fig. 2

In the proposed SV system with anti-spoofing mechanisms, the test utterance is verified as in standard SV systems. The speaker is considered an impostor if the SV system rejects the utterance. If the SV system accepts it as belonging to the target speaker, the verification is not considered final, undergoing a second stage of testing in which the utterance is fed to the anti-spoofing mechanism. If it is considered natural, the system accepts it as a target speaker utterance; otherwise it is rejected as a converted impostor utterance

The performance of the SV system with the anti-spoofing mechanism was reevaluated using the same metric (EER). This performance was also compared to that of a perfect anti-spoofing mechanism, which was simulated by assigning the correct output to each impostor trial. The performances of the SV system with real and ideal anti-spoofing mechanisms are shown in Table VI and the corresponding DET curves in Fig. 3.

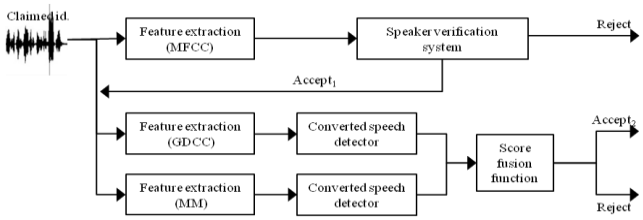

Fig. 2 SV system with anti-spoofing mechanism based on the fusion of two converted speech detectors

TABLE VI. EER % OF SV SYSTEM PERFORMANCE WITH REAL AND IDEAL ANTI-SPOOFING MECHANISM AGAINST NATURAL DATA; NATURAL DATA AND GMM-BASED IMPOSTORS; NATURAL DATA AND US BASED-IMPOSTORS

Voice conversion method	SV w/ real detectors (EER %)	SV w/ ideal detectors (EER %)
Baseline (no conversion)	10.3	9.4
GMM-based conversion	9.1	7.5
US-based conversion	8.9	7.4

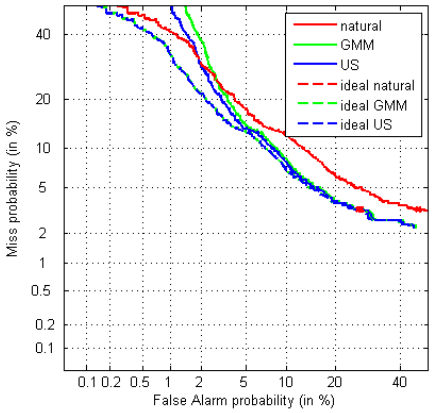

Fig. 3 DET curves of SV system performance with real and ideal anti-spoofing mechanism tested against natural data; natural data and converted GMM impostors; natural data and US impostors

We observe that the EER of the SV system with anti-spoofing mechanism drops as much as 11.1% compared to the EER presented on Table II. When comparing the SV system with real mechanism vs. ideal mechanism for converted speech detection, we observe that the 2.2% of GMM-based converted utterances corresponding to false positives accounted for an increase of the ideal EER of 1.6%; the 1.8% of false positives for the US-based conversion accounted for a 1.4% increase of the ideal EER. Finally, the 2.1% of false negatives of natural utterances corresponded to an increase of 0.9% of the ideal EER.

VI. CONLUSION

In this study, we evaluated the vulnerability of a state-of-the-art SV system against spoofing attacks by GMM-based and US-based converted speech. The experiment showed that the FA of the system deteriorated beyond what is acceptable for real life applications.

To manage the FA, we implemented state-of-the-art converted speech detectors based on features derived from the magnitude and GDFPS and compared their performances. Additionally we proposed a new more compact representation of the features and adopted a discriminative approach to model them. The proposed feature representation and modeling approach have outperformed the existing converted speech detectors.

We proceeded to fuse the scores of the best converted speech detectors and achieved better accuracy rates than with any standalone detector. That result strengthens the hypothesis that the short- and long-term features carried complementary information useful for converted speech detection. Finally we tested our SV system with anti-spoofing mechanism consisting of a fusion of two converted speech detectors using compact feature representation. We verified that the performance of the system was only marginally affected by the spoofing data.

ACKNOWLEDGMENT

The authors would like to thank Zhizheng Wu and Haizhou Li of the Nanyang Technological University, Singapore for providing the GMM-based and US-based converted speech corpora, and for their helpful suggestions. This work was supported by FCT grants PTDC/EIA-CCO/122542/2010 and PEst-OE/EEI/LA0021/2013, and EU Cost Action1206 De-identification for privacy protection in multimedia content..

REFERENCES

[1] J.P. Campbell Jr, “Speaker recognition: A tutorial,” Procedings of the IEEE, vol. 85, no. 9, pp, 1437-1462, 1997.

[2] D. Reynolds, “Automatic speaker recognition using gaussian mixture speaker models,” The Lincoln Laboratory Journal, vol. 8, no. 2, pp. 173-192, 1995.

[3] Y. Stylianou, O. Cappé, E. Moulines, "Continuous probabilistic transform for voice conversion," Speech and Audio Processing, IEEE Transactions on, vol. 6, no. 2, pp. 131-142, 1998.

[4] J.F. Bonastre, D. Matrouf, C. Fredouille, “Artificial impostor voice transformation effects on false acceptance rates,” Interspeech, 2007.

[5] Q. Jin, A. Toth, A.W. Black, and T. Schultz, “Is voice transformation a threat to speaker identification?,” Proc. ICASSP, pp. 4845–4848, 2008.

[6] Z. Wu, T. Kinnunen, E. Chng, H. Li, E. Ambikairajah, “A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case,” APSIPA ASC 2012.

[7] Z. Wu, E. Chng, H. Li, “Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition”, Interspeech 2012

[8] Z. Wu, X. Xiao, E. Chng, H. Li, “Synthetic speech detection using temporal modulation feature”, ICASSP 2013.

[9] L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, J. Yamagishi, “Detection of synthetic speech for the problem of imposture,” Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp.4844-4847, 2011

[10] F. Alegre, A. Amehraye, N. Evans, “Spoofing countermeasures to protect automatic speaker verification from voice conversion,” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on , pp.3068-3072, 2013

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transaction on Audio, Speech and Language Processsing, vol. 19, no. 4, 2011

[12] A. Kain, M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. Acoustics, Speech and Signal Processing, vol. 1, pp. 285–288, 1998.

[13] Y. Stylianou, O. Cappé, E. Moulines, "Statistical methods for voice quality transformation," EUROSPEECH, 1995.

[14] D. Sundermann, H. Hoge, A. Bonafonte,H. Ney, A. Black, S. Narayanan, “Text-Independent Voice Conversion Based on Unit Selection,” Acoustics, Speech and Signal Processing, 2006. ICASSP Proceedings, 2006. IEEE International Conference on , vol.1, pp.I-I, 2006

[15] J. Tribolet, "A new phase unwrapping algorithm," Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.25, no.2, pp.170-177, 1977

[16] B. Yegnanarayana, J. Sreekanth, A. Rangarajan, "Waveform estimation using group delay processing," Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.33, no.4, pp. 832-836, 1985

[17] L. D. Alsteris, K. Paliwal. 2007, “Short-time phase spectrum in speech processing: A review and some experimental results,” Digit. Signal Process, vol. 17, no. 3 pp. 578-616, 2007

[18] W. M. Campbell, D. E. Sturim, D. A. Reynolds, “Support vector machine using GMM supervectors for speaker verification”, IEEE Signal Processing Letters, vol. 13, no. 5, pp. 308-311, 2006

[19] “Bosaris toolkit [software package],” WWW page, March 2014, https://sites.google.com/site/bosaristoolki