Anti-spoofing: voice conversion

Nicholas Evans[*], Federico Alegre[†],
EURECOM, Biot, France
Zhizheng Wu[‡],
Nanyang Technological University (NTU), Singapore
Tomi Kinnunen[§],
University of Eastern Finland (UEF), Joensuu, Finland

Synonyms

Voice transformation; Speaker recognition; Speaker verification; Presentation attack

Definition

Voice conversion is a process which converts or transforms one speaker's voice towards that of another. The literature shows that voice conversion can be used to spoof or fool an automatic speaker verification system. State-of-the-art voice conversion algorithms can produce high-quality speech signals in real time and are capable of fooling both human listeners and automatic systems, including text-independent and text-dependent. Furthermore, since converted voice originates from a living person, traditional liveness detection countermeasures are not necessarily effective in detecting such attacks. With today's state-of-the-art algorithms producing high-quality speech with only few indicative processing artifacts, the detection of converted voice can be especially challenging.

Main Body Text

Introduction

Whereas the threat of spoofing to some biometric modalities has received considerable attention, there has been relatively little research to investigate vulnerabilities in the case of speaker recognition [1, 2, 3]. Early efforts focused

[*]evans@eurecom.fr
[†]alegre@eurecom.fr
[‡]wuzz@ntu.edu.sg
[§]tkinnu@cs.uef.fi

on impersonation and replay attacks. Impersonation is largely considered to be more effective in fooling human listeners rather than automatic recognition systems and the measuring of channel effects or audio forensic techniques can be used to detect replay attacks. More recent work has focused on high-technology attacks involving speech synthesis and voice conversion. The literature shows that the latter is particularly difficult to detect.

Voice conversion is a process which converts or transforms one speaker's voice towards that of another, specific target speaker. Conversion generally implies that the resulting speech 'sounds' like that of the target from a human-perception perspective, though some approaches convert only those aspects of speech which are most pertinent to automatic recognition, i.e. the spectral envelope. In this case, while the resulting speech may retain the prosodic qualities of the original speaker/impostor, it can be highly effective in overcoming automatic systems. With the capacity to produce high-quality convincing speech signals in real-time, today's state-of-the-art approaches to voice conversion present a potential threat to both text-dependent and text-independent systems.

Since they originate from a living person, traditional liveness detection countermeasures are not necessarily effective in detecting voice conversion attacks. Most countermeasures instead rely on the detection of specific processing artifacts. They require training examples in order to learn classifiers capable of detecting similarly treated, spoofed speech. In this sense countermeasures are specific to a particular voice conversion algorithm and are unlikely to generalise well to others.

This article overviews approaches to voice conversion, past work to assess the threat to automatic speaker verification and existing countermeasures.

Voice conversion

Several approaches to voice conversion were proposed in the 1980s and 1990s, e.g. [4, 5], and quickly spurred interests to assess the threat to automatic speaker verification (ASV), e.g. [6]. Voice conversion aims to convert or transform the voice of a source speaker ($\boldsymbol{X}$) towards that of a specific, target speaker ($\boldsymbol{Y}$) according to a conversion function $\mathcal{F}$ with conversion parameters $\boldsymbol{\theta}$:

$$\boldsymbol{Y} = \mathcal{F}(\boldsymbol{X}, \boldsymbol{\theta}).$$

The general process is illustrated in Figure 1. Most state-of-the-art ASV systems operate on estimates of the short-term spectral envelope. Accordingly, conversion parameters $\boldsymbol{\theta}$ are generally optimised at the feature level in order to maximise the potential for spoofing an ASV system which utilises the same or similar feature parameterisations.

While there is a plethora of different approaches to voice conversion in the literature, relatively few have been explored in the context of spoofing. The most common or influential among them are reviewed in the following.

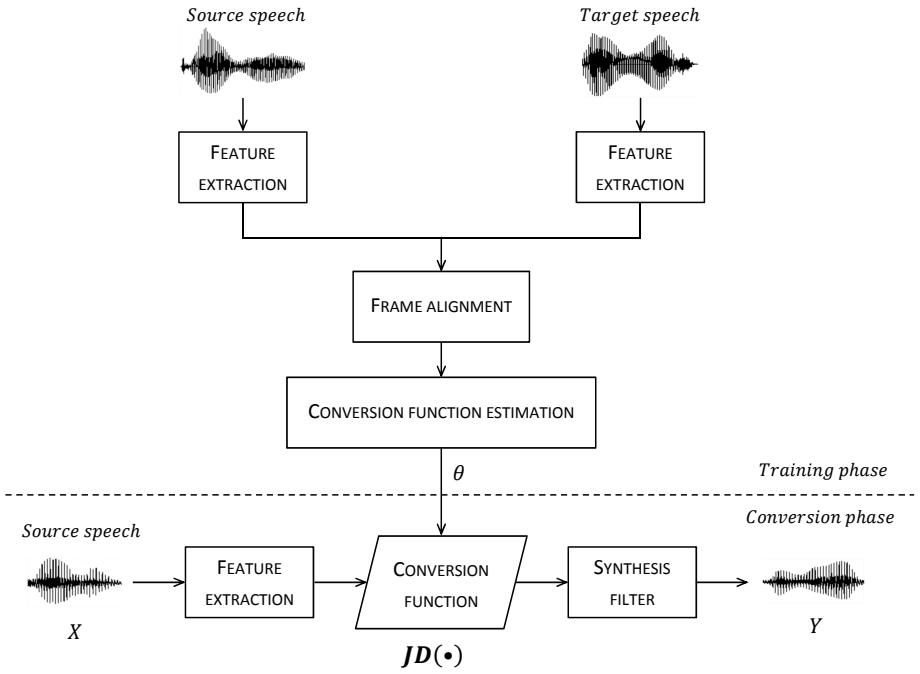

Figure 1: An illustration of general voice conversion using, e.g. joint density Gaussian mixture models (JD-GMMs). Figure adapted from [3].

Joint density Gaussian mixture models

As with most voice conversion approaches, and as illustrated in Figure 1, the popular *joint density Gaussian mixture model* (JD-GMM) algorithm [7] learns a conversion function using training data with a parallel corpus of frame-aligned pairs $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$. Frame alignment is usually achieved using dynamic time warping (DTW) on *parallel* source-target training utterances with identical text content. The combination of source and target vectors $\boldsymbol{z} = \left[\boldsymbol{x}^{\mathrm{T}} \boldsymbol{y}^{\mathrm{T}}\right]^{\mathrm{T}}$ is therefore used to estimate GMM parameters (component weights, mean vectors and covariance matrices) for the joint probability density of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The parameters of the JD-GMM are estimated using the classical expectation maximization (EM) algorithm in a maximum likelihood (ML) sense.

During the conversion phase, for each source speech feature vector $\boldsymbol{x}$, the joint density model is adopted to formulate a transformation function to predict the feature vector of the target speaker according to:

$$\mathcal{JD}\left(\boldsymbol{x}\right) = \sum_{l=1}^{L} p_l(\boldsymbol{x}) \left(\boldsymbol{\mu}_l^{(y)} + \boldsymbol{\Sigma}_l^{(xy)} \left(\boldsymbol{\Sigma}_l^{(xx)} \right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_l^{(x)} \right) \right)$$

where $p_l(\boldsymbol{x})$ is the posterior probability of the source vector $\boldsymbol{x}$ belonging to the l^{th} Gaussian. The trained conversion function is then applied to new source

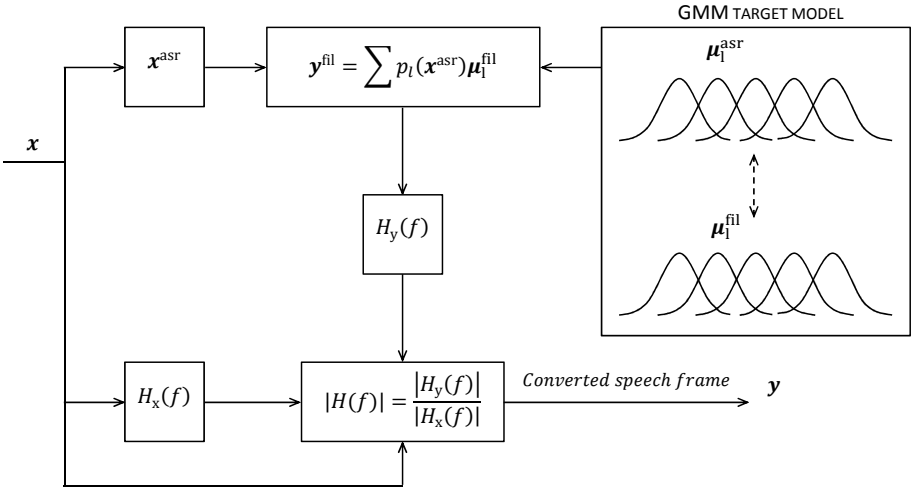

Figure 2: An illustration of Gaussian dependent filtering. Figure adapted with permission from [9].

utterances of arbitrary text content at run-time. In addition to parametric voice conversion techniques, *unit selection* – a technique which directly utilizes target speaker segments – is also effective in spoofing ASV [3, 8].

Gaussian dependent filtering

The work in [9] extends the concept of JD-GMM to utilise an explicit model of the target speaker at the core of the conversion process. It tests the vulnerabilities of ASV when the vocal tract information in the speech signal of a spoofer is converted towards that of the target speaker according to a Gaussian dependent filtering approach. As illustrated in Figure 2, the speech signal of a source speaker or spoofer, represented at the short-time frame level and in the spectral domain by $X(f)$, is filtered as follows:

$$\mathcal{GD}\left(X(f)\right) = \frac{|H_{\text{y}}(f)|}{|H_{\text{x}}(f)|}X(f)$$

where $H_{\text{y}}(f)$ and $H_{\text{x}}(f)$ are the vocal tract transfer functions of the target speaker and the spoofer respectively and $\mathcal{GD}\left(X(f)\right)$ denotes the result after voice conversion. As such, each frame of the spoofer's speech signal is mapped or converted towards the target speaker in a spectral envelope sense.

$H_{\text{y}}(f)$ is determined from a set of two GMMs. The first, denoted as the automatic speaker recognition (asr) model in the original work, is related to ASV feature space and utilized for the calculation of a posteriori probabilities whereas the second, denoted as the filtering (fil) model, is a tied model of linear predictive cepstral coding (LPCC) coefficients from which $H_{\text{y}}(f)$ is derived.

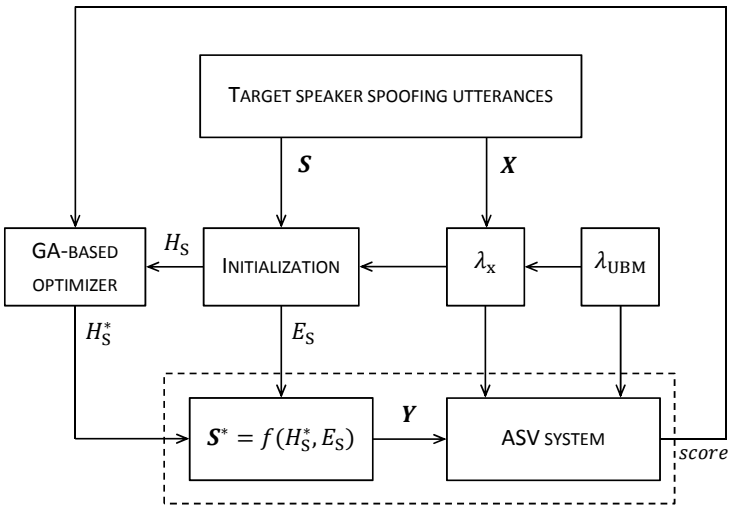

Figure 3: An illustration of artificial signal generation. Figure reproduced from [10].

LPCC filter parameters are estimated according to:

$$\boldsymbol{y}^{\text{fil}} = \sum_{l=1}^{L} p_l\left(\boldsymbol{x}^{\text{asr}}\right) \boldsymbol{\mu}_l^{\text{fil}}$$

where $p\left(\boldsymbol{x}^{\text{asr}}\right)$ is the posterior probability of the vector $\boldsymbol{x}^{\text{asr}}$ belonging to the l^{th} Gaussian in the asr model and $\boldsymbol{\mu}_l^{\text{fil}}$ is the mean of l^{th} Gaussian belonging to the fil model, which is tied to the l^{th} Gaussian in the asr model. $H_{\text{y}}(f)$ is estimated from $\boldsymbol{y}^{\text{fil}}$ using a LPCC to linear prediction (LP) coefficient conversion and a time-domain signal is synthesized from converted frames with a standard overlap-add technique. Resulting speech signals retain the prosodic aspects of the original speaker (spoofer) but reflect the spectral-envelope characteristics of the target while not exhibiting any perceivable artifacts indicative of manipulation. Full details can be found in [9].

Artificial signals

Spoofing with artificial signals [10] is an extension to the idea of voice conversion. Certain short intervals of converted voice yield particularly high scores or likelihoods. These short intervals can be further optimised and concatenated to produce arbitrary-length signals which reflect both the short-term static and dynamic characteristics of a target speaker. While resulting signals are not representative of intelligible speech, they are nonetheless effective in overcoming typical ASV systems which lack any form of speech quality assessment.

Let $\boldsymbol{S} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$ be a short sequence of consecutive speech frames selected from an utterance of the targeted speaker. As illustrated in Figure 3, the

algorithm seeks a new sequence of speech frames $\boldsymbol{S}^*$ which maximises the score of a given ASV system (which is assumed to represent the targeted system) and thus the potential for spoofing. Each frame $\boldsymbol{c}_t$ belonging to $\boldsymbol{S}$ is initially transformed in the frequency domain with voice conversion which gives:

$$\mathcal{AS}\left(C(f)\right) = \frac{|H_{\mathrm{c}}^*(f)|}{|H_{\mathrm{c}}(f)|}C(f)$$

While the set of excitations $E_{\mathrm{S}} = \{E_{\mathrm{c}_1}(f), E_{\mathrm{c}_2}(f), \ldots, E_{\mathrm{c}_n}(f)\}$ remains the same as the ones extracted from $\boldsymbol{S}$, optimisation is applied to identify a set of filters $H_{\mathrm{S}}^* = \{H_{\mathrm{c}_1}^*(f), H_{\mathrm{c}_2}^*(f), \ldots, H_{\mathrm{c}_n}^*(f)\}$. Instead of estimating each filter independently using Equation 1, however, the set of filters is jointly optimized using a genetic algorithm. Full details can be found in [10].

Spoofing

Reviewed below is some of the past work which has investigated ASV vulnerabilities to the specific approaches to voice conversion described above.

Even when trained using a non-parallel technique and telephony data, the baseline JD-GMM approach has been shown to increase significantly the false acceptance rate (FAR) of state-of-the-art ASV systems [11]. Even if speech so-treated can be detected by human listeners, experiments involving five different ASV systems showed universal susceptibility to spoofing. With a decision threshold set to the equal error rate (EER) operating point, the FAR of a joint factor analysis (JFA) system was shown to increase from 3% to over 17% whereas that of an i-vector probabilistic linear discriminant analysis (PLDA) system increases from 3% to 19%. The unit-selection approach was shown to be even more effective and increased the FARs to 33% and 41% for the JFA and PLDA systems respectively.

The work reported in [9] investigated vulnerabilities to voice conversion through the Gaussian dependent filtering of the spectral-envelope. Voice conversion was applied using the same feature parametrisations and classifier as the ASV system under attack. Results thus reflect the worst case scenario where an attacker has full knowledge of the recognition system and show that the EER of a GMM-based ASV system increases from 10% to over 60% when all impostor test samples were replaced with converted voice.

Experiments to assess vulnerabilities to artificial signals are reported in [10]. As illustrated in Figure 4, detection error trade-off (DET) profiles show that the EER of a standard GMM system increases from almost 10% to over 60% (1st and 3rd profiles respectively). That for a factor analysis (FA) system increases from 5% to almost 65%. Since artificial signals result from the further optimisation of small intervals of converted voice which attain a particularly high likelihood, it is perhaps not surprising that they provoke especially high error rates. Encouragingly, since artificial signals are entirely non-intelligible and non-speech-like, their detection is relatively straightforward, as discussed next.

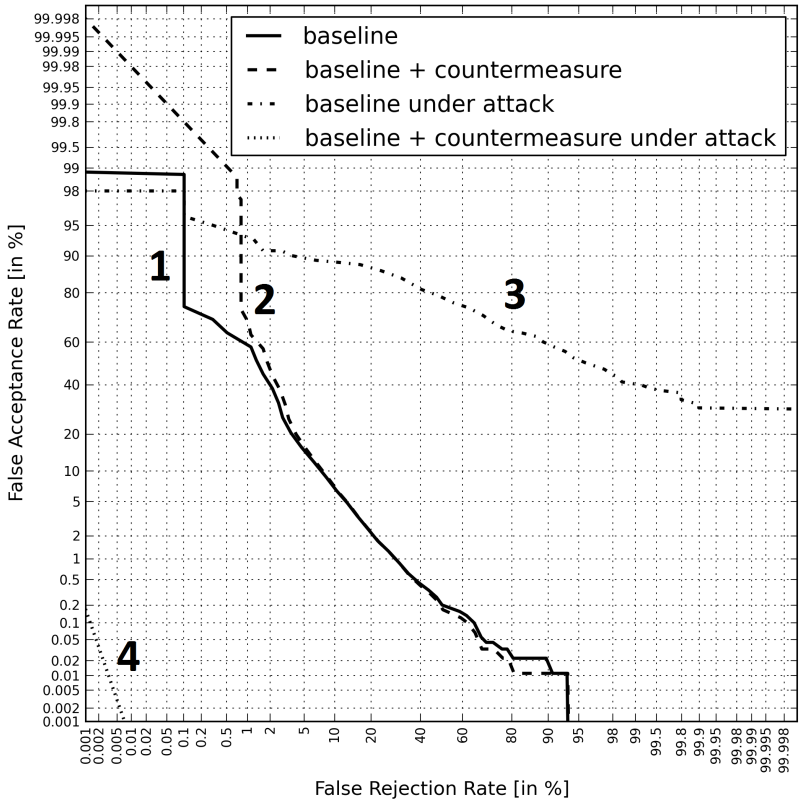

Figure 4: Example detection error trade-off profiles illustrating (i) the performance of a baseline GMM ASV system with naive impostors (ii) the same with active countermeasures, (iii) the baseline system where impostor accesses are replaced with artificial signal spoofing attacks and (iv) the same with active countermeasures. Profiles 2 and 4 correspond to a fixed countermeasure operating point where the threshold is tuned to give an FAR of 1%. Figure based on [10] and produced with the TABULA RASA Scoretoolkit: `http://publications.idiap.ch/downloads/reports/2012/Anjos_Idiap-Com-02-2012.pdf`.

Countermeasures

As the above shows, current ASV systems are essentially 'deaf' to conversion artifacts caused by imperfect signal analysis-synthesis models or poorly trained conversion functions. Tackling such weaknesses provides one obvious strategy to implement spoofing countermeasures.

Some of the first work to detect converted voice [12] draws on related work

in synthetic speech detection and considers phase-based countermeasures to JD-GMM and unit-selection approaches to voice conversion. The work investigated two different countermeasures, referred to as the cosine normalization and frequency derivative of the phase spectrum. Both countermeasures aim to detect the absence of natural speech phase, an artifact indicative of converted voice. The two countermeasures are effective in detecting converted voice with EERs as low as 6.0% and 2.4% respectively. In [11], the detector is combined with speaker verification systems for anti-spoofing. With a decision threshold set to the equal error rate (EER) operating point, baseline FARs of 3.1% and 2.9% for JFA and PLDA systems respectively fall to 0% for JD-GMM voice conversion attacks and to 1.6% and 1.7% for unit-selection attacks.

Phase-based countermeasures may be bypassed, however, by approaches to voice conversion which retain natural speech phase, i.e. approaches such as Gaussian-dependent filtering [9]. Noting that this approach to voice conversion produces speech signals of reduced short-term variability, the work reported in [13] investigated a countermeasure based on the average pair-wise distance between consecutive feature vectors. The approach captures greater levels of dynamic information beyond that in traditional features and is successful in detecting converted voice with real-speech phase with an EER of under 2.7%.

With a view to more generalised countermeasures, the work in [14] investigated the detection of converted voice and artificial signals using so-called local binary pattern (LBP) texture analysis of speech spectrograms. An utterance-level feature is used to detect the absence of natural, dynamic variability characteristic of genuine speech in a so-called textrogram. While performance is inferior to the approach proposed in [13], the countermeasure is less dependent on prior knowledge and successful in detecting different forms of spoofing.

Finally, a new approach to generalised countermeasures is reported in [15]. Extending the idea of LBP analysis to a one-class classification approach dependent only on training data of genuine speech, an SVM-based classifier is shown to give a detection EER of a little over 5% for converted voice, as illustrated in Figure 5. Even better results of 0.1% and 0% are obtained for speech synthesis and artificial signal attacks respectively. These results show the potential for generalised countermeasures, but also that converted voice is particularly difficult to detect. Countermeasure effects on the performance of the same GMM ASV system as in [10] are illustrated for artificial signal attacks in Figure 4. The second profile illustrates the effect on licit trials whereas the fourth profile illustrates the effect of spoofed trials. In both cases the countermeasure threshold is tuned to give a false reject rate (FRR) of 1%. First, for all but the lowest FRRs, the effect of the countermeasure on licit transactions is shown to be negligible. Second, for a fixed ASV FAR of 10%, the FAR is seen to fall from almost 90% to 0%. The effect of the same countermeasure on a state-of-the-art i-vector system is reported in [15].

In summary, while voice conversion is undoubtedly a high-technology attack beyond the means of the lay person, there is sufficient evidence that it presents a potential threat to the reliability of automatic speaker verification. Encouragingly, however, there is also significant progress to develop suitable

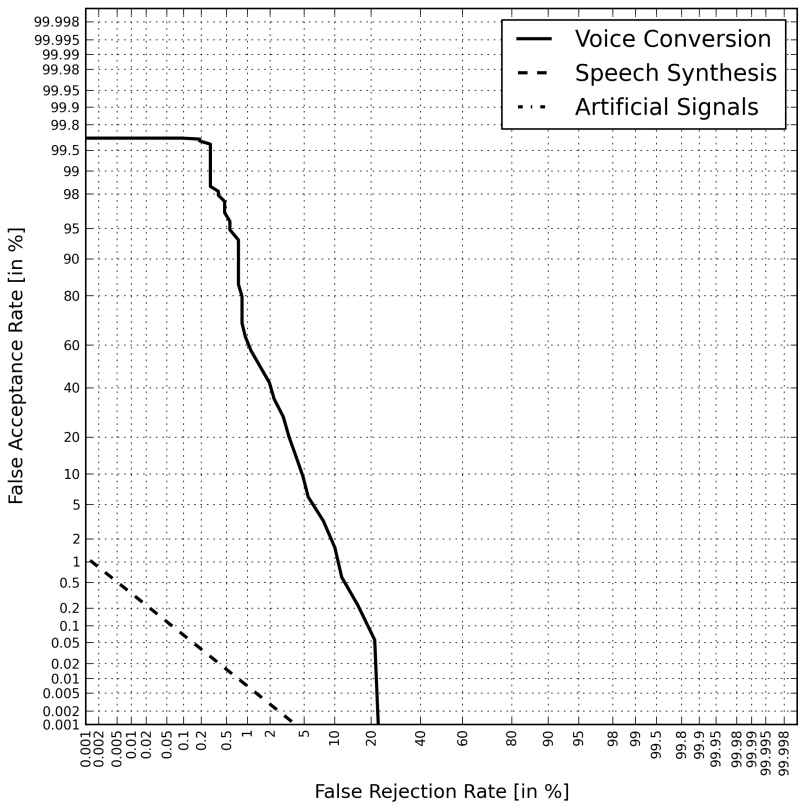

Figure 5: A detection error trade-off plot illustrating countermeasure performance independently from ASV. The profile for artificial signals is not visible since the EER is 0%. Figure reproduced from [15].

countermeasures and new initiatives to encourage research in the area [1]. In the future, standard datasets, protocols and metrics will be required so that effort can be focused on text-dependent scenarios [8] and generalised countermeasures capable of detecting unforeseen spoofing attacks [15]. Collaboration between voice conversion and automatic speaker verification researchers is also needed to ensure that systems are robust to state-of-the-art conversion algorithms.

Related Entries

References

[1] Evans, N., Kinnunen, T., Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification, INTERSPEECH, Proceedings of (2013)

[2] Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F., De Leon, P: Anti-spoofing for speaker recognition, chapter in Handbook of biometric anti-spoofing, Marcel, S., Li, S. Z., Nixon, M., Eds., Springer (2014)

[3] Wu, Z., Li, H.: Voice conversion and spoofing attack on speaker verification systems. Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Proceedings of (2013).

[4] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization, Acoustics, Speech, and Signal Processing (ICASSP), Proceedings of the 1988 IEEE International Conference on, v. 1, pp 655–658 (1988)

[5] Stylianou, Y., Cappé, O., Moulines, E.: Continuous probabilistic transform for voice conversion, Speech and Audio Processing, IEEE Transactions on, v. 6(2), pp. 131–142 (1998)

[6] Pellom, B. L., Hansen, J. H. L.: An experimental study of speaker verification sensitivity to computer voice-altered imposters, Acoustics, Speech, and Signal Processing (ICASSP), Proceedings of the 1999 IEEE International Conference on, v. 2, pp 837–840 (1999)

[7] Kain, A., Macon, M. W.: Spectral voice conversion for text-to-speech synthesis, Acoustics, Speech and Signal Processing (ICASSP), Proceedings of the 1998 IEEE International Conference on, pp. 285–288 (1998)

[8] Wu, Z., Larcher, A., Lee, K.A., Chng, E.S., Kinnunen, T., Li, H.: Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints. INTERSPEECH, Proceedings of (2013)

[9] Matrouf, D., Bonastre, J.-F., Fredouille, C.: Effect of speech transformation on impostor acceptance, Acoustics, Speech and Signal Processing (ICASSP), Proceedings of the 2006 IEEE International Conference on (2006)

[10] Alegre, F., Vipperla, R. Evans, N.: Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals. INTERSPSEECH, Proceedings of (2012)

[11] Wu, Z., Kinnunen, T., Chng, E.-S., Li, H., Ambikairajah, E.: A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Proceedings of (2012).

[12] Wu, Z., Chng, E. S., Li, H.: Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. INTERSPEECH, Proceedings of (2012)

[13] Alegre, F., Amehraye, A., Evans, N.: Spoofing countermeasures to protect automatic speaker verification from voice conversion. Acoustics, Speech and Signal Processing (ICASSP), Proceedings of the 2013 IEEE International Conference on (2013)

[14] Alegre, F., Vipperla, R., Amehraye, A., Evans, N.: A new speaker verification spoofing countermeasure based on local binary patterns, INTERSPEECH, Proceedings of (2013)

[15] Alegre, F., Amehraye, A., Evans, N.: A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns, Biometrics: Theory, Applications and Systems (BTAS), Proceedings of the International Conference on (2013)