Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 143–153

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Playback attack detection for text-dependent speaker verification over telephone channels ☆

Jakub Gałka *, Marcin Grzywacz, Rafał Samborski

AGH University of Science and Technology, Faculty of Computer Science, Electronics and Telecommunications, Poland

Received 21 May 2014; received in revised form 1 December 2014; accepted 3 December 2014
Available online 11 December 2014

Abstract

Playback attacks constitute one of the biggest threats in biometric speaker verification systems, in which a previously recorded passphrase is played back by an unprivileged person in order to gain access. This paper features a description of the playback attack detection (PAD) algorithm, designed to protect text-dependent speaker verification systems from the aforementioned spoofing attacks. The paper also describes the usage of spectral landmarks and score normalization methods in the playback detection algorithm. Different factors are discussed in terms of the performance of the algorithm. The authors investigate two issues: (1) extracting the PAD features which are robust against channel variations and (2) the robustness of the algorithm in adverse acoustical environments (e.g. office, street, cocktail party noise). The experiments are performed on a prepared speech corpus containing 4187 occurrences of a passphrase spoken by 175 speakers. The results of the experiment show the equal error rate (EER) to be as low as 1.0%. These findings demonstrate that such spoofing-oriented playback attacks can be effectively detected and should not be considered a significant argument against applications of text-dependent speaker verification.

Keywords: Speaker verification; Playback attack detection; Telephone channel; Spectral landmarks

1. Introduction

The task of biometric speaker verification is to accept or reject the identity claim of the speaker based on a sample of the speaker's voice. Telephone-based automatic speaker verification (by use of telephone channel) has already been a subject of research (Murthy et al., 1999; Kinnunen et al., 2012). Despite the fact that such systems perform very well, reaching relatively low EERs in demanding testing scenarios (NIST, 2012), consumers and organizations still have their doubts in the context of high-security applications (e.g. e-banking). One of the prevailing arguments against voice biometry concerns common passphrase text-dependent systems, in which the passphrase uttered by the speaker does not change from one login attempt to another. This enables the possibility of breaking into such systems by playing back a recording obtained earlier, using a microphone or any other eavesdropping method (e.g. malicious mobile software). This type of attack is called a playback attack and is available to anyone with minimal signal processing knowledge.

One of the solutions to this problem is to use a text-prompted system in which the user is asked to speak a randomly selected phrase for each access attempt. It is worth noting that such systems are more sensitive to other types of attacks (such as the concatenation of previously recorded digits) (Lindberg and Blomberg, 1999), and, due

☆ This work was supported by the Polish National Centre for Research and Development – Applied Research Program under Grant PBS1/B3/1/2012 titled "Biometric voice verification and identification".
* Corresponding author.
E-mail addresses: jgalka@agh.edu.pl (J. Gałka), mar.grzyw@gmail.com (M. Grzywacz), rafal.samborski@gmail.com (R. Samborski).

http://dx.doi.org/10.1016/j.specom.2014.12.003
0167-6393/

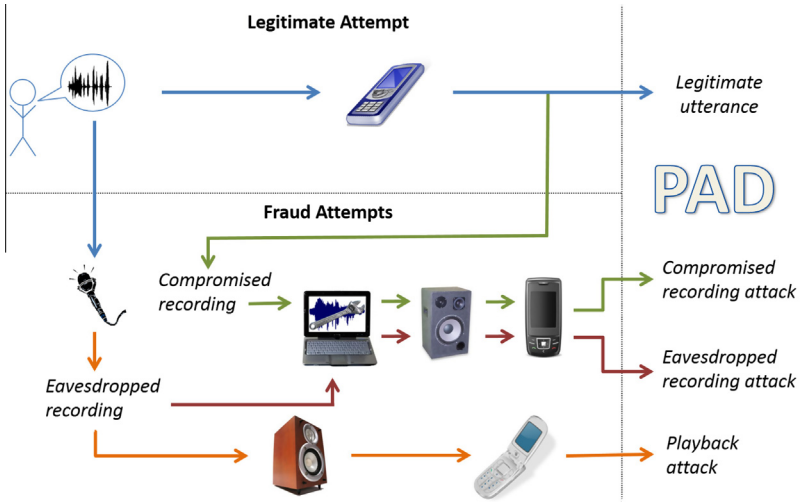

Fig. 1. Illustration of both legitimate and fraud verification attempts described in this document.

to the fact that the system cannot use lexical knowledge in its assessment, they achieve higher error rates, as compared to text-dependent solutions (Boves and den Os, 1998). The presented work focused on safeguarding text-dependent systems against playback attacks.

Several methods of playback attack detection are described, the bibliography on this subject not being very extensive. Employing direct spectral features such as the low frequency ratio was investigated in Villalba and Lleida (2011). The comparison of maps containing the highest peaks of the magnitude spectrum was described in Stevenson (2008). Another method, making use of a specific channel pattern, was presented in Wang et al. (2011). Despite the possibility of the normalization of similarity scores dramatically increasing the effectiveness of any of the aforementioned methods, there is little existing research on the subject. Shang and Stevenson (2010) successfully used the relative similarity score, which resulted in a reduction of the EER from 11.94% to 6.81%. None of the authors presented any kind of analysis of the impact of noise present in the attacking recording on detection performance. This is one of the reasons for the existence of the aforementioned algorithm. One of the objectives of this work was to achieve high noise robustness in a wide range of signal-to-noise ratio (SNR) of root mean square values. Another goal was taking advantage of the features available in devices which require high-speed data processing and have limited memory resources, such as embedded systems, physical biometric locks or other small-scale consumer electronics.

The method described in this paper uses both spectral features and score normalization to obtain a robust algorithm that addresses the issues of operating in an adverse acoustic environment, such as the one mentioned above. The paper is divided as follows: In Section 2, the core PAD algorithm is described, the corpus recorded for use in the experiments is described in Section 3, Section 4 provides the results of the conducted evaluations and covers the method of score normalization, in Section 5 conclusions are made and future work is discussed.

To improve the clarity of the paper, verification scenarios are presented in Fig. 1 and the following terms are defined:

Target: a privileged user, owner of data protected by biometric security.
Impostor: an unauthorized person claiming to be the owner of protected data, who attacks the system by modifying a previously acquired recording of the privileged (*Target*) user.
Legitimate: a non-playback-based verification attempt of the target.
Fraud: a playback attack by an impostor.
Authentic recording: a recording of a successful target verification attempt acquired server side.
Eavesdropped recording: a recording intercepted by an impostor on the client-side of the telecommunication channel during a legitimate verification attempt.
Compromised recording: a recording intercepted by an impostor on the server side of the telecommunication channel during a legitimate target's verification attempt, or a recording stolen from the server's user database.
Playback recording: the eavesdropped recording played back by the impostor and received at the server side of the telecommunication channel.

2. PAD algorithm

In this section, the PAD algorithm is presented. The concept of this solution is based on the music recognition system presented in Wang (2003) and Ellis (2009). Wang's idea of the algorithm is based on comparing recordings on the basis of the similarity of the local configuration of maxima pairs extracted from spectrograms of verified and reference recordings. According to the author of Wang (2003), the

algorithm is computationally efficient, as well as resistant to noise and channel distortions. These features are reproducible even in the presence of noise and Global System for Mobile communications (GSM) codec compression.

These statements, describing the operability of Wang's algorithm, show the possibility of using the algorithm to detect playback attacks in voice biometric systems. The utilized markers of each recording are unique and are presumed to be detectable in the processed copy, even in the presence of various types of distortion, including additive and convolutional noise, as well as degradation due to speech transmission coding. The computational efficiency of the algorithm and the high compression of the pattern of the recording are beneficial for the use of this algorithm in the context of PAD.

In this paper, particular properties and parameters of the aforementioned algorithm have been modified in order to better suit speech signals. Important alterations include the extraction of spectral landmarks, the scoring method, and the steps of classification. The core of the PAD algorithm consists of three major steps: (1) the preparation of the utterance spectrogram (2) spectral feature analysis and extraction of the local maxima pair matrix from the spectrogram (3) evaluating the similarity between the tested recording and the available set of reference recordings of the same speaker. A comparison is performed in order to decide whether the upcoming utterance is a fraud or a legitimate attempt.

The algorithm depends on the following operational parameters: (1) *peaks per frame* – the maximum number of spectral peaks accepted in one 64 ms time-frame (in this algorithm the maximum number is set to 5 peaks per frame, which guarantees both reasonable sensitivity and computational performance), (2) *pairs per peak* – the maximum number of pairs assigned to each peak in the *target region* (4 pairs per peak are used herein), (3) *target region* – the region, in which the spectral peaks are combined into pairs with a *seed peak*. The *seed peak* is the starting point of the *target region*, (4) *density* – the maximum number of accepted maxima pairs per second, (5) *spreading width* – the width (measured with frequency bins) of the decaying threshold surface assigned to each spectral peak in which the greater the value, the less peaks in the vicinity, (6) *decay rate* – the rate of the decay of the threshold surface following each peak, the value depending on density.

The values of parameters such as *density* and *pairs per peak* were increased, as the speech signal (as opposed to the music tracks) is characterized by greater variability in time and frequency in regard to signal duration. The algorithm, in its original form was prepared for longer signals (songs), while in the case of a biometric system, the speech signal lasts 1 to 3 s only. Thus, it was necessary to extract more spectral peaks per second and increase the incidence of the extracted landmarks, so that they appear in each significant phone of the uttered biometric password. Another important change was the correction of the *decay rate* parameter in order to adjust it to the modified settings of the landmark incidence mentioned earlier. All the settings discussed above were set to optimize *a posteriori* the accuracy of the algorithm during the development of the playback detector.

2.1. Spectrogram preparation

The first step of the PAD algorithm is to prepare a spectrogram of the upcoming speech utterance. 512-point Fast Fourier Transform (FFT) with 64 ms Hamming windowing and 32 ms overlapping is used to obtain the spectral features of the recording.

These values have been determined experimentally to provide the optimal spectral resolution for the extraction of the most significant features, and such settings produced the best results. Let the spectrogram S be a $N \times K$ matrix, where $S_{n,k}$ is a log-amplitude of the spectrogram at a given time frame n and frequency bin k and $K = 256$, which is equal to half the FFT length. To reduce the boundary effects introduced in the next steps of the procedure, the spectrogram is normalized by its mean spectral value

$$S^*_{n,k} = S_{n,k} - \mu_S, \tag{1}$$

where

$$\mu_S = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K} S_{n,k} \tag{2}$$

for $n = 1, \ldots, N$ and $k = 1, \ldots, K$.

A high-pass digital filter

$$H(z) = \frac{1 - z^{-1}}{1 - \alpha z^{-1}} \tag{3}$$

may be applied to each of the K spectrogram frequency bins along N consecutive time-frames in order to increase the contrast of the spectrogram, as well as improving maxima extraction efficiency. The closer the pole is to $z = 1$, the less spectral maxima are found in further processing steps. In the case of this study, the value $\alpha = 0.98$ was found to produce the best results during the performance test.

2.2. Maxima pair extraction

A simplified description of the local maxima pair extraction method is presented in this section. Details of the implementation can be found in Wang (2003) and Ellis (2009).

The set of spectral peak candidates is created by selecting elements of spectrogram S considered to be local maxima

$$(n_i, k_i) : S_{n,k} > \max\{S_{n,k-1}, S_{n,k+1}, d + S_{n-1,k}, S_{n+1,k}\}, \tag{4}$$

where d is a temporal peak-masking coefficient.

The obtained set of peak coordinates (n_i, k_i) is then pruned in order to eliminate insignificant peaks. That is, for each n-th frame, only the top-5 strongest spectral maxima are left in the set. All other frequency bins are set to

zero. The resulting sparse spectrogram is then processed further. The elimination of peak candidates is performed by way of thresholding selected candidates with the envelope $e_n(k)$, obtained by a convolution of the spectral frame with a Gaussian window. The width of the spreading Gaussian window can be adjusted during system evaluation. Only peak candidates higher than the envelope

$$(n_i, k_i) : S_{n,k} > e_n(k)/\rho, \tag{5}$$

where $\rho = 1.1$ is an overmasking factor, are eventually selected as the set of spectral maxima used for utterance parametrization. The value of ρ has been determined by experimentation. The described thresholding is similar to frequency masking, well known from perceptual audio processing methods.

Having the set of local maxima of the spectrogram, inside the *target region*

$$R_i(n,k) = \{(n,k) : n_i < n < n_i + r \ \wedge \ k_i - p < k < k_i + p\} \tag{6}$$

of each peak, the *seed* peak is then combined with other peaks from the target region into pairs. Parameters r and p determine the size of the *target region* along the time and frequency dimension respectively. The maximum number of pairs originating from each *seed* peak is limited by selecting the closest peak candidates. Examples of spectrograms with their maxima and maxima pairs are presented in Fig. 2. The figure presents an example of an authentic utterance and an example of a playback utterance. The general scheme of the peak-pair extraction is depicted in Fig. 3. The description of calculating the similarity between two recordings is given in Section 2.3.

The set of maxima pairs is represented by a 4-column matrix G, in which each row corresponds to one pair of

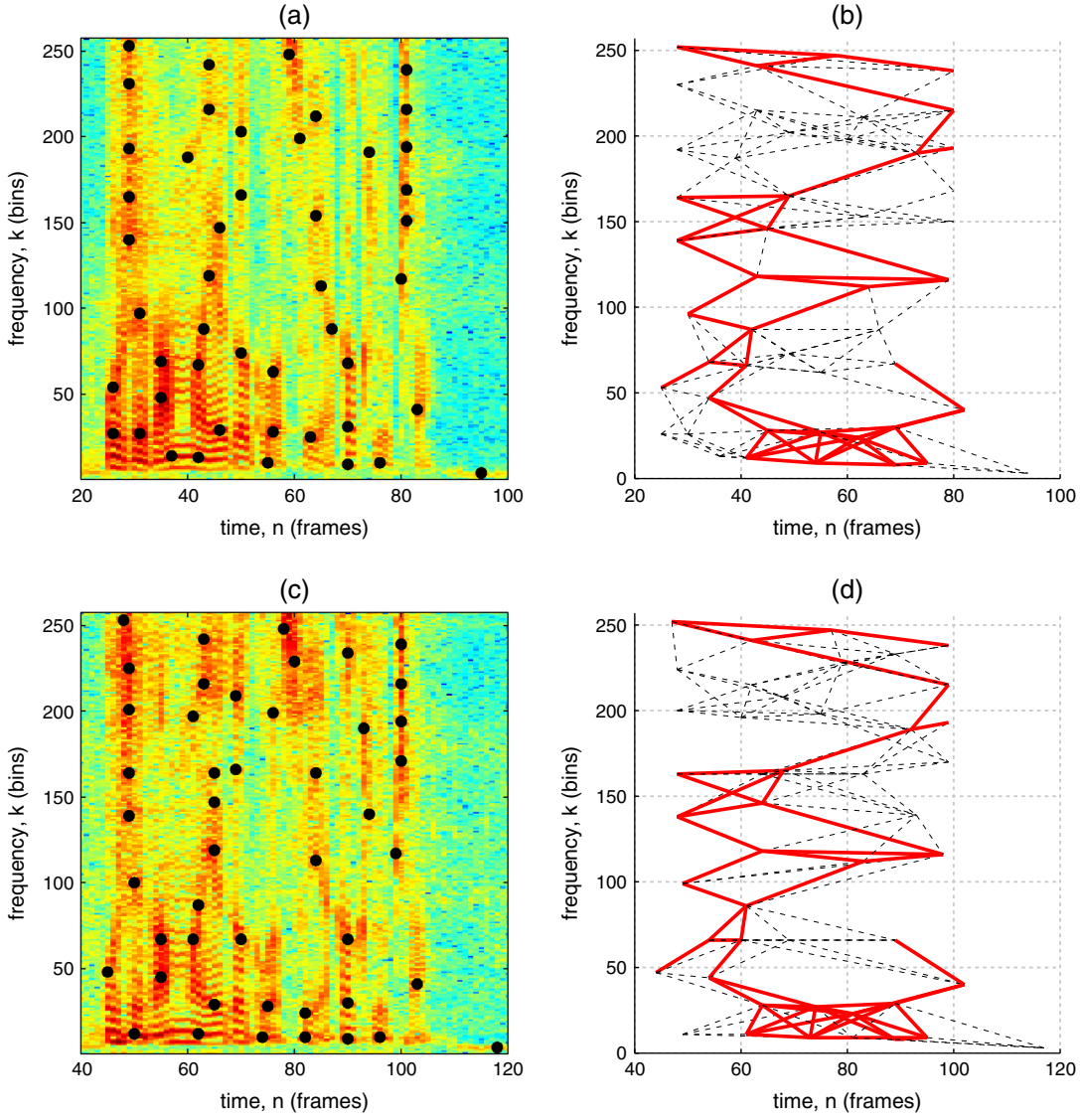

Fig. 2. Examples of authentic and playback utterances. Spectrograms with peaks: (a) authentic utterance, (c) playback utterance. Maxima pairs constellations: (b) authentic utterance, (d) playback utterance. Red lines show the similarity between an authentic recording and its playback version. Simplified features of utterance sets are shown to maintain the clarity of the figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

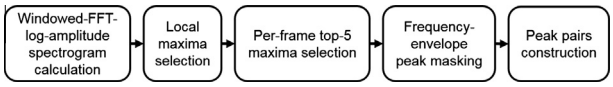

Fig. 3. Overview of the peak-pair extraction procedure.

maxima. Each row consists of 4 integer values: (k_i, n_i) – coordinates of the seed peak; (dk_i, dn_i) – frequency and time shifts between peaks paired with the seed peak. The matrix is sorted row-wise and duplicated rows are removed. The first redundant column is then removed as well. The matrix G of the original utterance is ready to be saved in the system's database and to be used as a reference in the scoring phase.

2.3. Similarity calculation

In this step, the pair-matrix G_{test} of the tested utterance is compared with all the pair-matrices of a particular target, stored for reference during previous verification attempts.

The comparison is conducted in such a way as to preserve the time-ordering of the maxima pairs. Each comparison results in the number of rows which are identical in both matrices. These numbers create the hits vector U of M length, where M is the number of reference recordings of a particular speaker in a database. In order to compute L, a set of baseline system scores, the ratio

$$L(m) = \frac{U(m)}{Y_m}, \tag{7}$$

of each element of U to the number of rows of its corresponding reference G_m is calculated, where Y_m is the number of rows in m-th reference matrix G_m. For example, if a comparison is made between tested utterance's pair-matrix G_{test} with reference pair matrix G_m of recording m from the speaker database, the ratio of the number of obtained identical rows to the number of all G_m rows is calculated. The PAD algorithm returns the value L_{max}, which is the maximum score of the baseline detection scores vector L. The obtained value is then used as a baseline similarity score to test the playback attack hypothesis. Baseline similarity score means a raw score without normalization.

3. Corpus

No proper speech corpus, containing both authentic and playback recordings, was available for comparative testing and benchmarking of the algorithm during the extensive research of the literature. Moreover, the collections which met the required criteria were not immediately available due to proprietary rights. Therefore, a new corpus was created to perform the necessary performance tests. For this purpose, recordings of 175 participants were collected, using a telephone channel. The voice samples were recorded in acoustically different environments by use of various models of mobile phones. Every participant repeated the passphrase utterance several times, which resulted in a total number of 4187 recordings. The passphrase was the Polish utterance "Używam mojego głosu jako klucza" (/uʒ i vammɔjɛgɔgwɔsujakɔkluʧâ/ in International Phonetic Alphabet (IPA) phonetic transcription), which can be literally translated as "I am using my voice as the key". The database of authentic recordings was recorded using Voice over IP (VoIP) technology employing a Gigaset C610 IP telephone using G.711, G.722, G.726, and G.729 codecs. All the recorded files where also checked using the Polish automatic speech recognition system (Ziółko et al., 2011), to confirm the passphrase correctness.

During the development of the algorithm, a subset of recordings was used in order to find the best values of the parameters listed in Section 2. This limitation was necessary due to the large number of parameters, the values of which had to be optimized, which in turn resulted in a large number of system evaluations. The development corpus consists of 324 authentic recordings and 324 playback recordings. The test sets of the recordings, used for the evaluation presented in Section 4, are listed at the beginning of each experiment's description.

3.1. Attack types

For 21 of the participants (11 female and 10 male), 81 of the recordings described above were also recorded with two additional microphones (see Fig. 4) to simulate eavesdropping playback attack preparation (eavesdropped recordings). The first of the two microphones (AKG C5) was located near (up to 10 cm) the mouth of the recorded person, the second one (Alphard ETP-280) was placed about 50 cm from the person. The sampling frequency for the telephone channel recordings was set to 8 kHz, with 16 bit resolution. The eavesdropped recordings were acquired using a sampling rate of 44.1 kHz, as well as 16 bit resolution. A Lexicon Lambda sound card was used for signal acquisition. Recordings simulating playback attacks were then created by playing back the eavesdropped recordings through a regular desktop computer speaker (Creative Inspire T10)

Fig. 4. Recording setup. The telephone channel (1) and both microphones ((2) near-field, (3) far-field) signals were recorded simultaneously with a multichannel PC sound card.

to the telephone's built-in microphone (Gigaset C610 IP telephone). Both the near-field and the far-field eavesdropped recordings were played back at two different distances from the microphone of the telephone. This resulted in 4 playback setups and a total number of 324 *playback* recordings which were used in the *Type 1 (Playback) attack*. Additionally, the eavesdropped recordings were convolved with an impulse response of the speaker-microphone set and encoded using a software-based Adaptive Multi Rate-Narrow Band GSM codec and VoIP channels to simulate playing them back to the system as *Type 2 (eavesdropping-based) attacks*. For these attacks, the method of eavesdropping the recording has the biggest impact on the final results. For 101 of the participants (38 female and 63 male), 1724 of the authentic recordings were modified in the same way in order to simulate situations, in which the impostor is capturing the signal on the server side of the telecommunication channel. Due to this, the recording becomes compromised, which is a *Type 3 (compromising-based) attack*. In this case, the biggest impact on the final result was the method of encoding the recording.

4. Results of the experiment

Two kinds of tests were conducted to assess the performance of PAD: (1) *legitimate* – when the recording is new, and has not been previously used by the user to gain access to the biometric system, and (2) *playback attack* – when the recording is a modified copy of one of the recordings that were previously used by the speaker for biometric verification. PAD may result in two different types of errors – false detection (FD), and missed detection (MD). A false detection error occurs when PAD claims that a legitimate (non-playback) attempt is an attack, while a missed detection error happens when PAD does not detect actual playback attacks. In the case of PAD, a missed detection error is much more dangerous, due to the chance of the playback attack succeeding, and the intruder gaining access to protected data. In the case of a false detection error, the user will be asked to repeat the password, and both the system and the data remain safe. Depending on the value of the decision threshold, we can estimate the values of the false detection rate (FDR) and missed detection rate (MDR). This notation was decided upon deliberately, as these errors should not be mistaken with the errors of the biometric part of the speaker verification system, which are usually called False Acceptance Rate (FAR) and False Rejection Rate (FRR).

As a measure of a system's performance, the evaluation can include (1) equal error rate *EER*, the value of the intersection of FDR and MDR curves, and (2) maximum accuracy Acc_{Max} of the system: the ratio of the maximal possible value of correctly classified attempts to the number of conducted trials. In order to examine the interdependence of FDR and MDR for various threshold values, detection error trade-off (DET) curves (Martin et al., 1997) were plotted in Figs. 6 and 7.

In the context of PAD, a thorough analysis of noise robustness and channel variations is very important for system security. Such analysis allows one to check the robustness of PAD to non-stationary distortion present in processed recordings. Such deformations of the signal always work for the benefit of the impostor, as any distortion may cause the attacking recording to be less similar to the authentic reference recording stored in the database. Due to this fact, the following paragraphs contain: (1) an experiment with varying normalization methods, (2) an experiment with varying attack types including the analysis of DET curves, (3) an experiment with channel variations – different location of playback devices, and (4) a noise resistance experiment.

4.1. Score normalization

Most decision-making systems based on detection-score thresholding gain accuracy and robustness when score normalization techniques are used. This was proven to be particularly important in speaker verification systems (Matsui and Furui, 1995, 1994; Ariyaeeinia and Sivakumaran, 1997; Ramos-Castro et al., 2007). Applying different score normalization methods to the PAD algorithm allows us to check how the algorithm works with each of them, and which are appropriate for implementation in this context. For the reader's convenience, brief definitions of normalization methods used for the evaluation of PAD are given below.

Let L_C be a so-called *cohort*-score vector of baseline detection scores L, where the maximum score L_{max} is excluded from the original score set. Normalization approaches based on L_C cohort scores have been described in Zigel and Cohen (2003). Several methods of cohort-based score normalization are mentioned, such as

$$L_{Cnorm_1} = \frac{L_{max}}{\max(L_C)}, \tag{8}$$

$$L_{Cnorm_2} = L_{max} - \mu_{L_C}, \tag{9}$$

$$L_{Cnorm_3} = \frac{L_{max}}{\mu_{L_C}}. \tag{10}$$

In the case of cohort normalization, to normalize scores using T-normalization (Barras and Gauvain, 2003) a cohort of L_C scores was used. The size of the L_C cohort is limited to 30 elements only (the closest to the L_{max} value). T-normalization was performed using the mean μ_T and standard deviation σ_T parameters of the $L_{C'}$ cohort:

$$L_{Tnorm} = \frac{L_{max} - \mu_T}{\sigma_T + \epsilon}, \tag{11}$$

where $\epsilon = 0.00431$ when $\sigma = 0$ is used as a backup, and $\epsilon = 0$ otherwise. The residual value of ϵ had been precalculated offline from test utterances (both *legitimate* and *playback attacks*), where $\sigma \neq 0$, and the value of ϵ reflects the mean of all non-zero σ values from such tests.

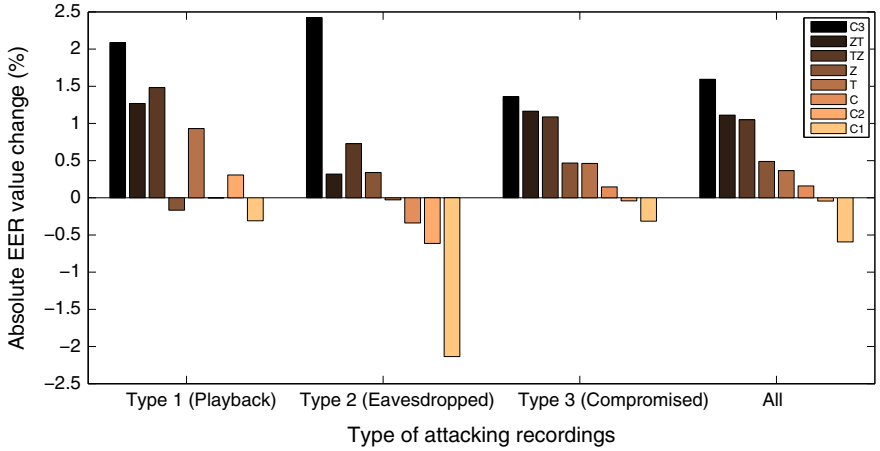

Fig. 5. Influence of various normalization methods on EER values for different types of attacking recordings. Reference *baselines* are $EER_{Type1(Playback)} = 0.6\%, EER_{Type2(Eavesdropped)} = 2.2\%, EER_{Type3(Compromised)} = 0.9\%, EER_{All} = 1.0\%$. Term *baselines* means EER values before score normalization.

Another well-known method of score normalization is Z-normalization (Ramos-Castro et al., 2007), where the mean μ_Z and standard deviation σ_Z parameters, calculated for each of the speaker's reference recordings, were obtained by scoring the recording against a set of randomly chosen reference recordings of other speakers:

$$L_{Znorm} = \frac{L_{max} - \mu_Z}{\sigma_Z + \epsilon}. \quad (12)$$

The main advantage of the Z-norm is that the parameters μ_Z and σ_Z can be pre-calculated offline.

Combinations of different normalization methods were also investigated. Such methods include ZT-norm (conducting T-normalization first, followed by Z-normalization), TZ-norm and the normalization described by the formula

$$L_{Xnorm} = \frac{L_{max} - \frac{\mu_T + \mu_Z}{2}}{\frac{1}{4}\sqrt{\sigma_T^2 + \sigma_Z^2}}, \quad (13)$$

which combines the properties of T- and Z-normalization simultaneously.

The experiment containing varying normalization methods involved 2372 attack attempts: (1) 324 playback utterances *Type 1 (Playback)*, (2) 324 post-processed eavesdropped utterances *Type 2*, (3) 1724 modified authentic utterances *Type 3 (Compromising-based)* and 226714 legitimate attempts. The presented normalization methods, as well as the aforementioned types of attacking recordings were evaluated under different testing conditions, both played back by the impostor, as well as modified algorithmically. These types of attacking recordings differ in quality, the amount non-linear distortions (channel variations), and coding methods.

The absolute percentage value changes in the obtained equal error rates (EER) of the PAD system for different normalization methods, against the baseline $EER_{Type1(Playback)} = 0.6\%, EER_{Type2(Eavesdropped)} = 2.2\%, EER_{Type3(Compromised)} = 0.9\%, EER_{All} = 1.0\%$, are presented in Fig. 5. Negative values indicate lower EER and better system performance.

The relative EER changes in Fig. 5 show that many of the tested normalization methods are not efficient in the case of our PAD system. For *Type 1* recordings, most of the normalization methods worsen the EER value. Only C_1norm and *Znorm* improved the performance, as the EER was lower by 0.3% and 0.2% respectively. For *Type 2* recordings, three of the normalization methods lowered the EER, the best one being C_1norm, which lowered the EER by 2.1%, although with C_2norm or *Xnorm*, the EER value dropped by about 0.5%. For *Type 3* recordings and for all recording types altogether, C_1norm gave the best results of all the normalization methods. These facts influenced our choice of C_1norm as the best normalization method for the PAD algorithm.

4.2. Varying attack types

The second experiment was performed using three types of attack recordings, described in Section 3. The aim was to test the baseline effectiveness of the system against different types of playback attacks. The experiment involved 2372 attack attempts: (1) 324 playback utterances *Type 1 (Playback)*, (2) 324 post-processed eavesdropped utterances *Type 2*, (3) 1724 modified authentic utterances *Type 3 (Compromised)*, and 226714 legitimate attempts. The results of the experiment with varying attack types are presented in (Table 1). In the table, aside from the system performance baseline, the normalization methods were limited to three due to the large number of tested normalization methods. To assess system performance, EER and Maximum Accuracy values are presented for each type of attack. The best of the obtained values are in bold. The table shows that the EER is relatively low, which confirms that the implemented system is accurate. The fact that EER = **0.6%** is the lowest value for the recordings *Type 1 (Playback)*, which best reflect real attack attempts, confirms that PAD works very well.

Table 1
The performance of PAD on various types of attacks against the baseline system and the three most efficient normalization methods for each class of recordings (Fig. 5). Best performance is in bold.

Playback	Acc_{Max} (%)	*EER* (%)	Eavesdropped	Acc_{Max} (%)	*EER* (%)
Xnorm	99.56	1.2	*Xnorm*	98.91	3.7
$Cnorm_1$	**99.92**	**0.6**	$Cnorm_1$	**99.65**	**2.2**
$Cnorm_2$	99.34	0.9	$Cnorm_2$	98.71	4.0
Baseline	99.28	0.9	*Baseline*	98.65	4.3
Compromised	Acc_{Max} (%)	*EER* (%)	All	Acc_{Max} (%)	*EER* (%)
Xnorm	99.76	1.3	*Xnorm*	99.71	1.7
$Cnorm_1$	99.78	**0.9**	$Cnorm_1$	**99.80**	**1.0**
$Cnorm_2$	**99.80**	1.1	$Cnorm_2$	99.75	1.5
Baseline	99.78	1.2	*Baseline*	99.73	1.6

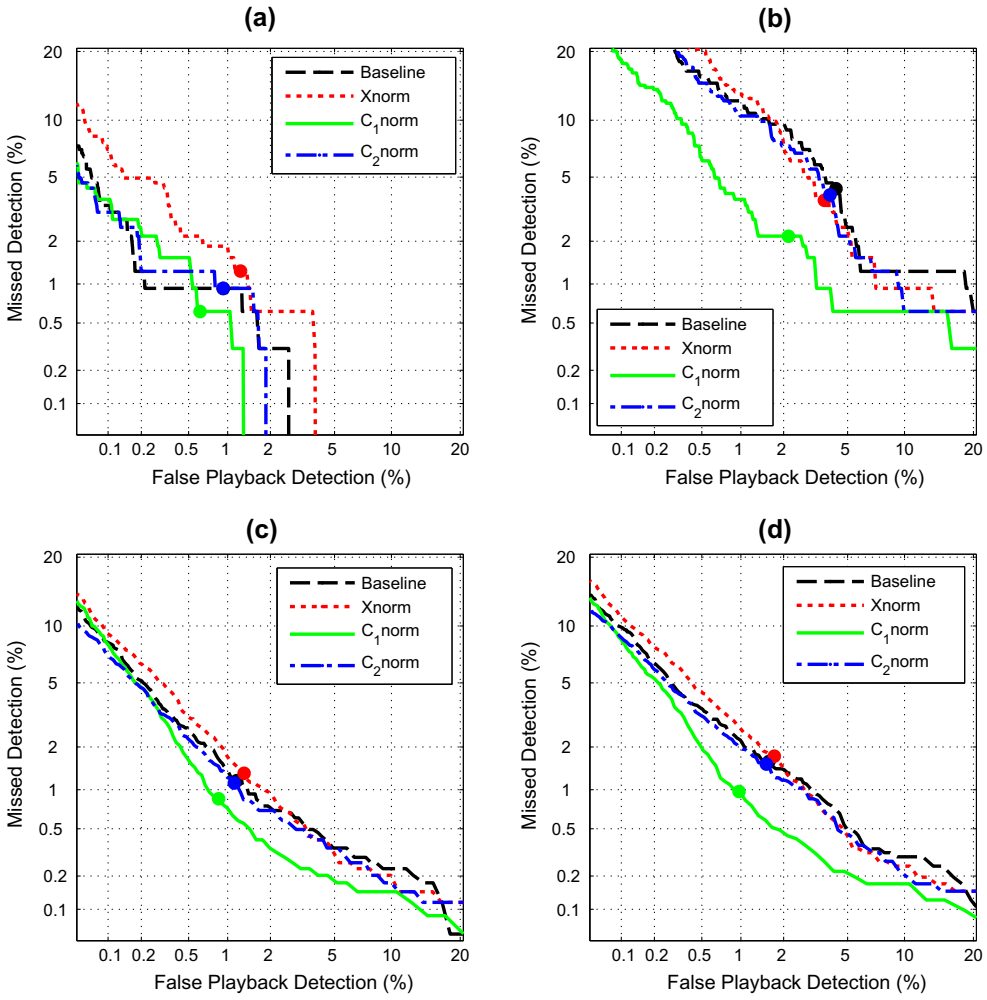

Fig. 6. DET curves for: (a) playback recording attacks, (b) eavesdropped recording attacks, (c) compromised recording attacks, (d) average for all types of attacks. Circles mean operating points of EER. The roughness of the curves for low values of missed detections is discussed in Section 4.2

To show the detection error trade-off between FDR and MDR values, DET (Fig. 6) curves are presented for each of the groups of attacks against the baseline system, as well as the three most efficient normalization methods. It is worth mentioning that rough shapes of DET curves are not necessarily due to a small testing base, but the fact that a discrete process is being observed (finite, and no non-integer number of identical pairs can exist, such as 0, 1 or 3 for dissimilar recordings; only small, repetitive numbers).

In the case of a high security system scenario, in which PAD does not miss any playback attacks, the operating point of the system should be set to the value of 0% MDR. This point is reachable with FDR greater than 40% (see Fig. 6d). Such system settings have no practical

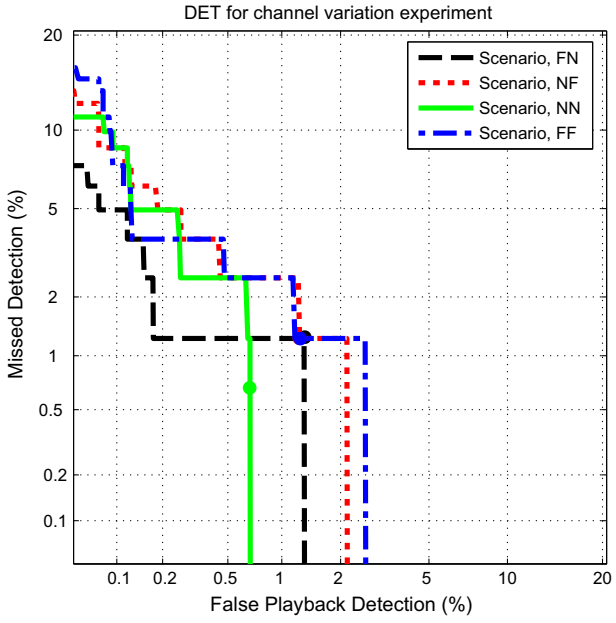

Fig. 7. DET curves for various scenarios of attacking recording *Type 1 (Playback)*. Scenario NN means near field location of impostor's microphone and loudspeaker, NF means near field location of impostor's microphone and far field location of impostor's loudspeaker, etc. Circles mean operating points of EER. The roughness of the curves was discussed in Section 4.2.

sense, as too many legitimate attempts would be denied. A better option is e.g. to reduce MDR to a value lower than 1%, with the FDR value not being significantly higher.

4.3. Channel variations

An experiment was conducted to verify the effectiveness of the system against a variety of channel distortion, caused by different locations of the eavesdropping microphone and a loudspeaker replaying the utterance. The experiment involved 324 attack attempts of *Type 1 (Playback)*, as well as all of the possible legitimate attempts. Attacking recordings of *Type 1 (Playback)* were separated into four classes, in regards to the different locations of the eavesdropping microphone acquiring the recording and the loudspeaker playing it back to the system. The first playback scenario occurred when both devices were located in the near field (*Scenario NN*). Another scenario occurred when the microphone was located in the near field and the loudspeaker was located in the far field (*Scenario NF*). *Scenario FN* represents the opposite situation, in which the microphone was located in far field and the loudspeaker was located in the near field. *Scenario FF*, the lowest quality scenario, occurred when both devices were situated in the far field. For each of these scenarios, DET curves were plotted in Fig. 7. The angularity of the DET curves is not necessarily due to a small base, but a discrete process, as mentioned above.

The best quality frauds (Scenario NN) were the easiest to detect, as for this class EER equals about 0.6%. For the medium quality recordings class (Scenarios NF, FN) and for the lowest quality recordings (Scenario FF) the value of EER equals around 1.1%. These values are acceptable, but the possibility of changing the decision threshold value in order to obtain low MDR should be taken into consideration. The results of the experiment show that the system is resistant to channel variation, depending on location of the eavesdropping devices.

4.4. Noise robustness

The last experiment was carried out in order to check PAD's resistance to non-stationary noise corruption of incoming playback recordings. The performance of PAD is presented over three different types of noise as *EER* and Acc_{Max} values in the function of different SNR. During

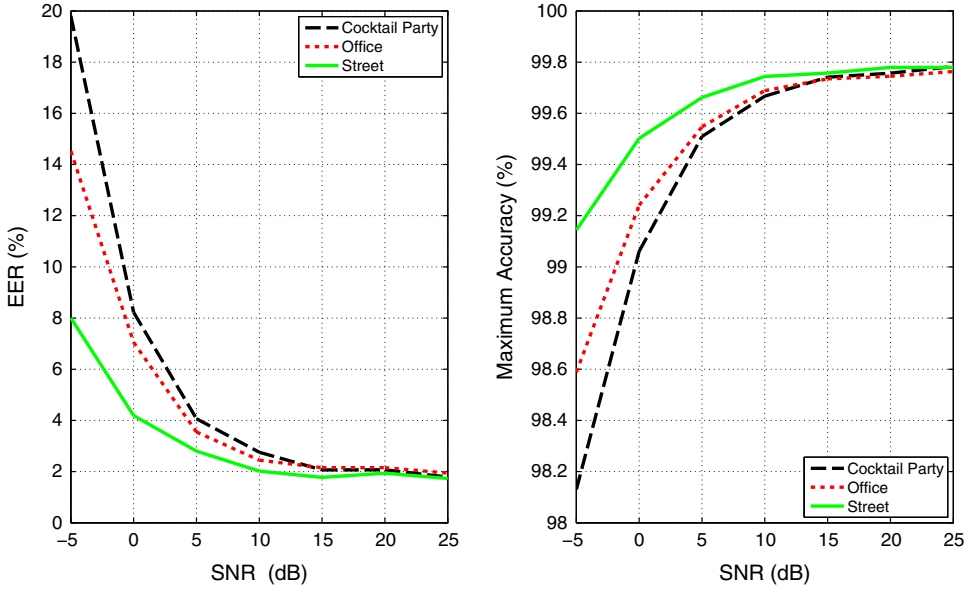

Fig. 8. The performance of PAD for recordings corrupted by three different types of noise with different SNR for scores normalized with C_2norm.

this experiment, the scores were normalized with $Cnorm_1$. In order to corrupt playback recordings, three kinds of non-stationary noise signals were used: (1) cocktail-party noise, (2) office noise, (3) street noise; with SNR ranging from −5 dB to 25 dB with a 5 dB step. The results are presented in Fig. 8. Plotted curves show that the presented system is robust to attacking recordings corrupted by noise with SNR greater than 10 dB. An EER value lower than 3% and a Maximum Accuracy value greater than 99% are both sufficient in the case of PAD, since the lower the SNR value, the smaller the chance of detecting an attack, but at the same time, the bigger the possibility that during the following stages, the speaker verification system will not accept such a low-quality fraud. The interchangeability of the tasks of these two systems is a very important phenomenon.

5. Summary

In this paper, the PAD system was presented. The idea behind this solution is based on the music recognition system proposed in Wang (2003) and Ellis (2009). PAD compares recordings on the basis of the similarity of robust spectral landmarks, which are reproducible and resistant to noise and channel variations.

The experiments evaluating the performance of PAD involved three attack types, which were described in detail in Section 3. An average EER value of 1.0% shows that the efficiency of the algorithm is high, and that the algorithm is able to detect most playback attacks. The experiment with channel variations, as well as the presence of noise proves that the system can deal with distorted speech signals. The experiment with varying normalization methods shows that, among well-known normalization methods, the performance of PAD is improved with the usage of C_1norm score normalization.

During the development of the system, two issues occurred, which need to be addressed in further research. One of them is the time and frequency modulation of speech signals. The algorithm itself is not robust enough to such signal modification, and may be cheated by attacks created with the use of pitch-shifting or time-scaling algorithms. In this case, the cooperation of biometric speaker verification and playback detection systems is very important, as the greater the signal modification rate, the greater the possibility that the recording would be rejected during the next stages of biometric speaker verification. This issue can also be addressed by distortion-mimicking pre-processing (analysis-by-synthesis approach). Expected forms of signal modification can be simulated prior to the execution of the PAD algorithm, allowing successful attack detection. Another issue is that in the context of system performance, the number of recordings of each of the users is more important than the number of users in the database. This affects score normalization and the computational efficiency of the algorithm.

Comparing the results of the PAD system to the expectations of the market, it is safe to assume the viability of using such a solution in real-world applications. This method, due to its localized spectral-pair-based detection, can be used (to some extent) to detect playback utterances prepared (tailored) artificially with the help of concatenation or more advanced speech modification methods. However, this issue should be studied in more depth in the future.

The remaining question is, in what way does the lexical content of the passphrase affect the performance of PAD. The authors plan to address this issue by evaluating the algorithm using the recently published RSR2015 speech database (Larcher et al., 2014). Another speech database, containing passphrases of various lexical content, is being prepared by the authors as well. Both speech data resources will be used to assess PAD's accuracy in more general lexical conditions.

This article does not cover the subject of the detection of playback attacks in full. Future work on the algorithm, as well as testing, will certainly include expanding the base of attacking recordings to present the effectiveness of the system with even greater reliability. The issues of time-scaling and pitch-shifting signal modification (including non-linear modification), as well as passphrase tailoring, will be studied as well.

References

Ariyaeeinia, A.M., Sivakumaran, P., 1997. Analysis and comparison of score normalisation methods for text-dependent speaker verification. In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (Eds.), EUROSPEECH, ISCA.

Barras, C., Gauvain, J., 2003. Feature and score normalization for speaker verification of cellular data. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), vol. 2, pp. II–49–52. doi:http://dx.doi.org/10.1109/ICASSP.2003.1202291.

Boves, L., den Os, E., 1998. Speaker recognition in telecom applications. In: 1998 IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications, 1998. IVTTA '98. Proceedings, pp. 203–208. doi:http://dx.doi.org/10.1109/IVTTA.1998.727721.

Ellis, D., 2009. Robust landmark-based audio fingerprinting. <http://labrosa.ee.columbia.edu/matlab/fingerprint/>.

Kinnunen, T., Wu, Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H. 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4401–4404. doi:http://dx.doi.org/10.1109/ICASSP.2012.6288895.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2014. Text-dependent speaker verification: classifiers, databases and rsr2015. Speech Commun. 60, 56–77.

Lindberg, J., Blomberg, M., 1999. Vulnerability in speaker verification – a study of technical impostor techniques. In: EUROSPEECH'99, pp. 1211–1214.

Martin, A.F., Doddington, G.R., Kamm, T., Ordowski, M., Przybocki, M.A., 1997. The det curve in assessment of detection task performance. In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (Eds.), EUROSPEECH, ISCA, pp. 1895–1898.

Matsui, T., Furui, S., 1994. Similarity normalization method for speaker verification based on a posteriori probability. ESCA Workshop Automat Speaker Recogn. Ident. Verif., 59–62

Matsui, T., Furui, S., 1995. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. Speech Commun. 17 (1–2), 109–116.

Murthy, H., Beaufays, F., Heck, L., Weintraub, M., 1999. Robust text-independent speaker identification over telephone channels. IEEE Trans. Speech Audio Process. 7 (5), 554–568. http://dx.doi.org/10.1109/89.784108.

2012 NIST speaker recognition evaluation results page. <http://www.nist.gov/itl/iad/mig/sre12results.cfm>.

Ramos-Castro, D., Firrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2007. Speaker verification using speaker- and test-dependent fast score normalization. Pattern Recogn. Lett. 28 (1), 90–98.

Shang, W., Stevenson, M., 2010. Score normalization in playback attack detection. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 1678–1681. http://dx.doi.org/10.1109/ICASSP.2010.5495503.

Stevenson, M., 2008. A playback attack detector for speaker verification systems. 2008 3rd International Symposium on Communications, Control and Signal Processing. IEEE, pp. 1144–1149. http://dx.doi.org/10.1109/ISCCSP.2008.4537397.

Villalba, J., Lleida, E., 2011. Preventing replay attacks on speaker verification systems. 2011 Carnahan Conference on Security Technology. IEEE, pp. 1–8. http://dx.doi.org/10.1109/CCST.2011.6095943.

Wang, A.L.-C., 2003. An industrial-strength audio search algorithm. In: Choudhury, S. and Manus, S. (Eds.), ISMIR 2003, 4th Symposium Conference on Music Information Retrieval, 2003, pp. 7–13, The International Society for Music Information Retrieval. http://www.ismir.net: ISMIR, October, pp. available: <http://www.ee.columbia.edu/dpwe/papers/Wang03-shazam.pdf>.

Wang, Z.-F., Wei, G., He, Q.-H., 2001. Channel pattern noise based playback attack detection algorithm for speaker recognition. In: 2011 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 4, pp. 1708–1713. doi:http://dx.doi.org/10.1109/ICMLC.2011.6016982.

Zigel, Y., Cohen, A., 2003. On cohort selection for speaker verification. In: INTERSPEECH, ISCA, pp. 2977–2980.

Ziółko, M., Gałka, J., Ziółko, B., Jadczyk, T., Skurzok, D., Masior, M., 2011. Automatic speech recognition system dedicated for polish. In: INTERSPEECH, ISCA, pp. 3315–3316.