Long-Term Spectral Statistics for Voice Presentation Attack Detection

Hannah Muckenhirn, *Student Member, IEEE,* Pavel Korshunov, *Member, IEEE,* Mathew Magimai.-Doss, *Member, IEEE,* and Sébastien Marcel, *Member, IEEE*

***Abstract*—Automatic speaker verification systems can be spoofed through recorded, synthetic or voice converted speech of target speakers. To make these systems practically viable, the detection of such attacks, referred to as presentation attacks, is of paramount interest. In that direction, this paper investigates two aspects: (a) a novel approach to detect presentation attacks where, unlike conventional approaches, no speech signal modeling related assumptions are made, rather the attacks are detected by computing first order and second order spectral statistics and feeding them to a classifier, and (b) generalization of the presentation attack detection systems across databases. Our investigations on ASVspoof 2015 challenge database and AVspoof database show that, when compared to the approaches based on conventional short-term spectral features, the proposed approach with a linear discriminative classifier yields a better system, irrespective of whether the spoofed signal is replayed to the microphone or is directly injected into the system software process. Cross-database investigations show that neither the short-term spectral processing based approaches nor the proposed approach yield systems which are able to generalize across databases or methods of attack. Thus, revealing the difficulty of the problem and the need for further resources and research.**

***Index Terms*—Presentation attack detection, anti-spoofing, spectral statistics, cross-database**

I. Introduction

THE goal of an automatic speaker verification (ASV) system is to verify a person through her/his voice. The system receives as input a speech sample along with an identity claim. It outputs a binary decision: the speech sample corresponds to the claimed identity or not. ASV systems can make two types of errors: reject a true or genuine claim referred to as false rejection, or accept a false or impostor claim referred to as false acceptance. ASV systems can be applied in different scenarios such as forensic or personal authentication. Although the ultimate goal is to have a system that is error free, in practice, the ASV systems are error prone and, depending upon the application, a trade-off between the error types exist. For example, in forensic applications, false rejections would be considered more costly, while in speech-based personal authentication applications, false acceptances would be considered more costly. This paper is concerned with an up-and-coming issue related to ASV systems in the latter scenario, i.e., personal authentication scenario.

Like any biometric system, ASV-based authentication systems can be attacked at different points [1], as illustrated in Fig. 1. In this paper, our interest lies in attacks at point (1) and point (2), called spoofing attacks, where the system can be attacked by presenting a spoofed signal as input. It has been shown that ASV systems are vulnerable to such elaborated attacks [2], [3]. As for points of attack (3) - (9), the attacker needs to be aware of the computing system as well as the operational details of the biometric system. Preventing or countering such attacks is more related to cyber-security, and is thus out of the scope of the present paper.

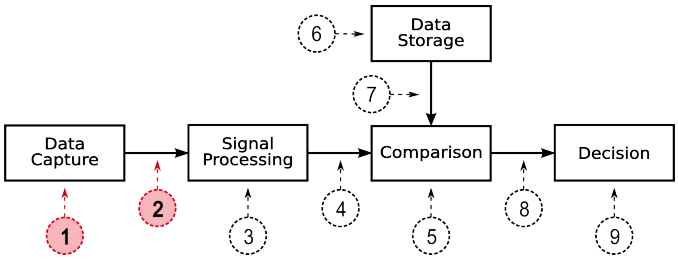


Fig. 1. *Potential points of attack in a biometric system, as defined in the ISO-standard 30107-1 [4]. Points 1 and 2 correspond respectively to attacks performed via physical and via logical access.*

Attack at point (1) is referred to as *presentation attack* as per ISO-standard 30107-1 [4] or as *physical access attack*. Formally, it refers to the case where falsified or altered samples are presented to the biometric sensor (microphone in the case of ASV system) to induce illegitimate acceptance. Attack at point (2) is referred to as *logical access attack* where the sensor is bypassed and the spoofed signal is directly injected into the ASV system process. The main difference between these two kinds of attacks is that in the case of physical access attacks, the attacker, apart from having access to the sensor, needs less expertise or little knowledge about the underlying software. Whilst in the case of logical access attacks, the attacker needs the skills to hack into the system as well as knowledge of the underlying software process. In that respect, physical access attacks are more likely or practically feasible than logical access attacks. Despite the technical differences, in an abstract sense this paper treats physical access attacks and logical access attacks as presentation attacks, as both are related to presentation of falsified or altered signal as input to the ASV system.

There are three prominent methods through which these attacks can be carried out, namely, (a) recording and replaying the target speakers speech, (b) synthesizing speech that carries target speaker characteristics, and (c) applying voice conversion methods to convert impostor speech into target speaker

Authors are with Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland, e-mail: firstname.lastname@idiap.ch

H. Muckenhirn is also with Ecole Polytechnique Fédérale de Lausanne, Switzerland

speech. Among these three, replay attack is the most viable attack, as the attacker mainly needs a recording and playback device. In the literature, it has been found that ASV systems, while immune to "zero-effort" impostor claims and mimicry attacks [5], are vulnerable to such elaborated attacks [2]. The vulnerability could arise due to the fact that ASV systems are inherently built to handle undesirable variabilities. The spoofed speech can exhibit variabilities that ASV systems are robust to and thus, can pass undetected.

As a consequence, developing countermeasures to detect presentation attacks is of paramount interest, and is constantly gaining interest in the speech community [3]. In that regard, the emphasis until now has been on logical access attacks, largely thanks to the "Automatic Speaker Verification Spoofing and Countermeasures Challenge" [6], which provided a large benchmark corpus containing voice conversion based and speech synthesis based attacks. As discussed in more detail in Section II, in the literature, countermeasure development has largely focused on investigating short-term speech processing based features that can aid in discriminating genuine speech from spoofed signal. This includes cepstral-based features, phase information, and fundamental frequency based information, to name a few.

The present paper focuses on two broad inter-connected research problems concerned with presentation attack detection (PAD), namely,

1) Most of the countermeasures developed until now have been built on top of standard short-term speech processing techniques that enable decomposition of speech signal into source and system, and develop countermeasures focusing on either one of them or both. However, both genuine accesses and presentation attacks are speech signals that carry the same high level information, such as message, speaker identity, and information about environment. There is little prior knowledge that can guide us to differentiate genuine access speech from presentation attack speech. Hence, a question that arises is: do we still need to follow standard short-term speech processing techniques for PAD? Aiming to answer this question, we develop a novel approach that does not make speech signal modeling related assumptions, such as quasi-stationarity. It simply computes the first and the second order spectral statistics over Fourier magnitude spectrum to detect presentation attacks without any dimensionality reduction.
2) As mentioned earlier, research on detecting presentation attacks has mainly focused on logical access attacks, though physical access attacks are more likely or practically easier. So a first set of questions that arises is: are physical access attack detection and logical access attack detection different? Would the methods developed for logical access attack detection be scalable to physical access attack detection? Towards that, we present benchmarking experiments on AVspoof corpus, which contains physical access attacks. Specifically, we use the recent work by Sahidullah *et al.* [7], which benchmarked several anti-spoofing systems for logical access attacks, as a starting point. We select several well-performing methods and evaluate them along with the proposed approach of using spectral statistics based features on physical access attack detection, through an open source implementation based on the Bob framework [8], and contrast them w.r.t. logical access attack detection. We then, in one of the first efforts, further study these aspects from cross-database and cross-attack perspective.

It is worth mentioning that a part of the results presented in the paper has appeared in [9] and in [10]. Specifically, the previous work on long-term spectral statistics (LTSS) [9] was limited to comparison to the top five systems of ASVspoof Interspeech 2015 challenge. In this paper, we study the LTSS-based approach in comparison to short-term spectral feature based systems selected from [7] and benchmarked in [10] on both ASVspoof and AVspoof databases. We also investigate the LTSS-based approach in a cross database scenario. Furthermore, we analyze different aspects related to the proposed LTSS-based approach, namely, (a) is it better to model the raw log magnitude spectrum, as done in previous works [7], [11], [12], or use statistics of the raw log magnitude spectrum, as done in the proposed approach? (b) Impact of the window size on the detection of physical access attacks and logical access attacks, (c) analysis of the system at decision level to get insight about the generalization capabilities in cross database studies, and (d) analysis of the trained models to understand the discriminative information learned by the LDA classifier for different types of attacks, including the importance of first order and second order statistics.

The remainder of the paper is organized as follows. Section II provides a background on the countermeasures developed for logical access attacks. Section III then motivates and presents the proposed spectral statistic based approach for PAD. Section IV presents the experimental setup. Section V presents the results and Section VI presents an analysis of the proposed approach and results obtained. Finally, in Section VII, we conclude.

II. Related work

As mentioned earlier, various methods have been proposed in the context of logical access attack detection. All these approaches can be broadly seen as development of a binary classification system. This involves extraction of features based on conventional short-term speech processing and training a classifier. In this section, we provide a brief overview about the methods. For a more comprehensive survey, please refer to [3], [13].

A. Features

In the literature, different feature representations based on short-term spectrum have been proposed for synthetic speech detection. These features can be grouped as follows:

1) magnitude spectrum based features with temporal derivatives [14], [15]: this includes standard cepstral features (e.g., mel frequency cepstral coefficients, perceptual linear prediction cepstral coefficients, linear prediction cepstral coefficients), spectral flux-based features

that represent changes in power spectrum on frame-to-frame basis, sub-band spectral centroid based features, and shifted delta coefficients.
2) phase spectrum based features [14], [16]: this includes group delay-based features, cosine-phase function, and relative phase shift.
3) spectral-temporal features: this includes modulation spectrum [16], frequency domain linear prediction [7], extraction of local binary patterns in the cepstral domain [17], [18], and spectrogram based features [19].

The magnitude spectrum based features and phase spectrum based features have been investigated individually as well as in combination [20], [21], [22], [23]. All the aforementioned features are based on short-term processing. However, some features such as modulation spectrum or frequency domain linear prediction tend to model phonetic structure related long-term information.

In addition to these spectral-based features, features based on pitch frequency patterns have been proposed [24], [25]. There are also methods that aim to extract "pop-noise" related information that is indicative of the breathing effect inherent in normal human speech [26].

B. Classifiers

Choosing a reliable classifier is especially important given possibly unpredictable nature of attacks in a practical system, since it is unknown what kind of attack the perpetrator may use when spoofing the verification system. Different classification methods have been investigated in conjunction with the above described features such as logistic regression, support vector machine (SVM) [7], [17], artificial neural networks (ANNs) [27], [11], and Gaussian mixture models (GMMs) [7], [14], [16], [20], [21], [22], [23]. The choice of classifier is also dictated by factors like dimensionality of features and characteristics of features. For example, in [7], GMMs were able to model sufficiently well the de-correlated spectral-based features of dimension 20-60 and yield highly competitive systems. Whilst in [12], ANNs were used to model large dimensional heterogeneous features.

The classifiers are trained in a supervised manner, i.e., the training data is labeled in terms of genuine accesses and attacks. During recognition or detection, the classifier outputs a frame level evidence or scores for each class, which are then combined to make a final decision. For instance, in the case of GMM-based classifier, the log-likelihood ratio is computed similarly to a Gaussian Mixture Model-Universal Background Model (GMM-UBM) ASV system, and is then compared to a preset threshold to make the final decision.

Leveraging on recent findings in machine learning, deep ANNs have also been employed to learn automatically the features using intermediate representations as input such as log-scale spectrograms [28] or filterbanks [27], [29], [30].

III. Proposed approach: long-term spectral statistics

This section first motivates the scientific basis for the use of long-term spectral statistics based information for PAD, and then presents the details of the proposed approach.

A. Motivation

In presentation attack detection, we face a situation where we need to discriminate a speech signal (genuine) against another speech signal (attack) without a good prior knowledge about the characteristics that distinguishes the two speech signals. In the literature, as discussed in Section II, approaches have been developed by applying conventional speech modeling techniques to extract features and then classify them. The difficulty stems from the fact that conventional speech modeling is equally applicable to both genuine access signals and attack signals, even when synthesized. More precisely, the synthesis and voice conversion systems are largely built around the notion of source-system modeling, which is also used for extracting features for PAD. Success of such approaches largely depends upon the details involved in source-system modeling, and consequently, may need more than a single feature representation. For instance, the top five systems in ASVspoof Challenge 2015 employed multiple features. In this paper, we take an approach where we make minimal assumptions about the signal. More precisely, we assume that the two signals have two different statistical characteristics, irrespective of what is spoken and who has spoken. One such statistical property is the means and variances of the energy distributed in the different frequency bins.

first order statistics: Long-term average spectrum (LTAS) is a first order spectral statistics that can be estimated either by performing a single Fourier transform of the whole utterance or by averaging the spectrum computed by windowing the speech signal over the utterance [31], [32]. Originally, the interest in estimating LTAS emerged from the studies on speech transmission [33] and the studies on intelligibility of speech sounds, specifically measurement of articulation index, which represents the proportion of average speech signal that is audible to a human subject [34]. Later in the literature, LTAS has been extensively used to study voice characteristics [32]. It is employed for example for the early detection of voice pathology [35] or Parkinson disease [36], or for evaluating the effect of speech therapy or surgery on the voice quality [37]. In addition to assessing voice quality, LTAS has also been used to differentiate between speakers gender [38] and speakers age [39], to study singers and actors voices [40], [41] and also to perform speaker verification [31]. First order statistics is interesting for developing countermeasures for presentation attacks as natural speech and synthetic speech differ in terms of both intelligibility and quality. In particular, during estimation of LTAS the short-term variation due to phonetic structures get averaged out, and thus facilitates study of voice source [32]. Modeling effectively voice source in statistical parametric speech synthesis systems is still an open challenge [42], [43]. This aspect can be potentially exploited to detect attacks by using LTAS as features.

second order statistics: Speech is a non-stationary signal. The energy in each frequency bin changes over the time. Natural speech and synthetic speech can differ in terms of such dynamics. Indeed one of the successful approaches to classify natural and synthetic speech signals is use of dynamic temporal derivative information of short-term spectrum as

opposed to static information [7]. Variance of magnitude spectrum can be seen as a gross estimate of such dynamics. More precisely, standard deviation is indicative of the dynamic range of the magnitude in a frequency bin. Thus, variance could be useful for detecting attacks.

Speech signal is acquired through a sensor, which has its own channel characteristics. Information about the channel characteristics can be modeled through spectral statistics. State-of-the-art speech and speaker recognition systems employ the first order spectral statistics, e.g. mean of cepstral coefficients[1] [46] and the second order spectral statistics, e.g. variance of cepstral coefficients to make the system robust to channel variability. Channel information, however, is a desirable information for the detection of both physical access attacks and logical access attacks. In the case of physical access attacks, the spoofed signal is played through a loud speaker, which is captured via the system microphone. Such channel effects are cues for detecting attacks. For instance, hypothetically should the channel effect of the recording sensor and the loud speaker be "perfectly" removed then detecting record-and-replay attack is a non-trivial task. Channel information is also interesting for detecting logical access attacks, as the spoofed speech signal obtained from speech synthesis or voice conversion systems is injected into the system, while the genuine speech signal is captured through the sensor of the system. In the literature it has been shown that first order and second order spectral statistics can be used to predict speech quality or quality assessment [47], [48]. In the case of both physical access attacks and logical access attacks, we can expect the speech quality to differ w.r.t the genuine speech signal.

The simplest approach to make minimal assumptions about the signal is to use the raw log-magnitude spectrum directly as feature input to the classifier. In that direction, the use of the short-term raw log-magnitude spectrum has been investigated in several works [7], [11], [12]. However, it has been found to perform poorly when compared to standard features such as Mel-frequency cepstral coefficients (MFCC). A potential reason for that can be that short-term raw log-magnitude spectrum contains several types of information, such as message, speaker, channel and environment. As we shall see later in Section VI-A, this puts onus on the classification method to learn the information that discriminates genuine access and attack. On the contrary, as explained above, the long term spectral statistics average out phonetic structure information [32], [49] and are indicative of voice quality as well as speech quality. Thus, we hypothesize that statistics of raw log magnitude spectrum can be effectively modeled for PAD when compared to raw log magnitude spectrum. The following section presents our approach in detail.

B. Approach

The approach consists of three main steps:

1) *Fourier magnitude spectrum computation*: the input utterance or speech signal x is split into M frames using a frame size of w_l samples and a frame shift of w_s samples. We first pre-emphasize each frame to enhance the high frequency components, and then compute the N-point discrete Fourier transform (DFT) $\mathcal{F}$, i.e., for frame m, $m \in \{1 \cdots M\}$:
$$X_m[k] = \mathcal{F}(x_m[n]), \tag{1}$$
where $n = 0 \cdots N-1$, with $N = 2^{\lceil \log_2(w_l) \rceil}$, and $k = 0 \cdots \frac{N}{2} - 1$, since the signal is symmetric around $\frac{N}{2}$ in the frequency domain. If $|X_m[k]| < 1$, we floor it to 1, i.e., we set $|X_m[k]| = 1$ so that the log spectrum is always positive. For each frame m, this process yields a vector of DFT coefficients $\mathbf{X}_m = [X_m[0] \cdots X_m[k] \cdots X_m[\frac{N}{2} - 1]]^{\mathrm{T}}$.
The number of frequency bins depends upon the frame size w_l as $N = 2^{\lceil \log_2(w_l) \rceil}$. In our approach, it is a hyper-parameter that is determined based on the performance obtained on the development set.
2) *Estimation of utterance level first order (mean) and second order (variance) statistics per Fourier frequency bin*: given the sequence of DFT coefficient vectors $\{\mathbf{X}_1, \cdots \mathbf{X}_m, \cdots \mathbf{X}_M\}$, we compute the mean $\mu[k]$ and the standard deviation $\sigma[k]$ over the M frames of the log magnitude of the DFT coefficients:
$$\mu[k] = \frac{1}{M} \sum_{m=1}^{M} \log |X_m[k]|, \tag{2}$$
$$\sigma^2[k] = \frac{1}{M} \sum_{m=1}^{M} (\log |X_m[k]| - \mu[k])^2, \tag{3}$$
$k = 0 \cdots \frac{N}{2} - 1$.
The mean and standard deviation are concatenated, which yields a single vector representation for each utterance.
3) *Classification*: the single vector long-term spectral statistic representation of the input signal is fed into a binary classifier to decide if the utterance is a genuine sample or an attack. In the present work, we investigate two discriminative classifiers: a linear classifier based on linear discriminant analysis (LDA) and a multi-layer perceptron (MLP) with one hidden layer.

IV. Experimental setup

We describe the details of the experimental setup in this section. All the systems described here are based on the open-source toolbox Bob[2] [8] and on Quicknet [50] and are reproducible[3].

A. Databases

We present experiments on two databases: (a) the automatic speaker verification spoofing (ASVspoof) database, which contains only logical access attacks; and (b) the audio-visual spoofing (AVspoof) database, which contains both logical and physical access attacks.

[1]Formally, the cepstrum is the Fourier transform of the log magnitude spectrum [44], [45].

[2]https://www.idiap.ch/software/bob/

[3]Source code: https://pypi.python.org/pypi/bob.paper.taslp_2017

1) ASVspoof: The ASVspoof[4] database contains genuine and spoofed samples from 45 male and 61 female speakers. This database contains only speech synthesis and voice conversion attacks produced via logical access, i.e., they are directly injected in the system. The attacks in this database were generated with 10 different speech synthesis and voice conversion algorithms. Only 5 types of attacks are in the training and development set (S1 to S5), while 10 types are in the evaluation set (S1 to S10). This allows to evaluate the systems on known and unknown attacks. The full description of the database and the evaluation protocol are given in [6]. This database was used for the ASVspoof 2015 Challenge and is a good basis for system comparison as several systems have already been tested on it.

2) AVspoof: The AVspoof database[5] contains replay attacks, as well as speech synthesis and voice conversion attacks both produced via logical and physical access. This database contains the recordings of 31 male and 13 female participants divided into four sessions. Each session is recorded in different environments and different setups. For each session, there are three types of speech:

- Reading: pre-defined sentences read by the participants,
- Pass-phrase: short prompts,
- Free speech: the participants talk freely for 3 to 10 minutes.

For physical access attack scenario, the attacks are played with four different loudspeakers: the loudspeakers of the laptop used for the ASV system, external high-quality loudspeakers, the loudspeakers of a Samsung Galaxy S4 and the loudspeakers of an iPhone 3GS. For the replay attacks, the original samples are recorded with: the microphone of the ASV system, a good-quality microphone AT2020USB+, the microphone of a Samsung Galaxy S4 and the microphone of an iPhone 3GS. The use of diverse devices for physical access attacks enables the database to be more realistic. This database is a subset of the one used for the BTAS challenge [51]. The training and development sets are the same while some additional attacks were recorded for the BTAS challenge in order to have "unknown" attacks in the evaluation set. Here, the types of attacks are the same in the three sets.

B. Evaluation Protocol

In both databases, the dataset is divided into three subsets, each containing a set of non-overlapping speakers: the training set, the development set and the evaluation set. The number of speakers and utterances corresponding to these three subsets are presented in Table I and in Table II respectively for the ASVspoof database and the AVspoof database.

The evaluation measure used in ASVspoof 2015 Challenge was equal error rate (EER), where the decision threshold τ_* is set as:

$$\tau_* = \arg\min_{\tau} |\text{FAR}_\tau - \text{FRR}_\tau|$$

More specifically, in both the development and evaluation set, the threshold is fixed independently for each type of attack with the EER criterion. Then, the performance of the system is evaluated by averaging the EER over the known attacks (S1-S5), the unknown attacks (S6-S10) and all the attacks.

[4]http://dx.doi.org/10.7488/ds/298

[5]https://www.idiap.ch/dataset/avspoof

TABLE I
NUMBER OF SPEAKERS AND UTTERANCES FOR EACH SET OF THE ASVSPOOF DATABASE: TRAINING, DEVELOPMENT AND EVALUATION.

data set	speakers		utterances	
	male	female	genuine	LA attacks
train	10	15	3750	12625
development	15	20	3497	49875
evaluation	20	26	9404	184000

TABLE II
NUMBER OF SPEAKERS AND UTTERANCES FOR EACH SET OF THE AVSPOOF DATABASE: TRAINING, DEVELOPMENT AND EVALUATION.

data set	speakers		utterances		
	male	female	genuine	PA attacks	LA attacks
train	10	4	4973	38580	17890
development	10	4	4995	38580	17890
evaluation	11	5	5576	43320	20060

In realistic applications, the decision threshold is a hyper-parameter that has to be set a priori. So evaluation of systems simply based on EER, where the optimal threshold is found on the evaluation set, may not reflect the realistic scenario well. A more realistic evaluation approach would be to determine τ_* on the development set and compute the half total error rate (HTER) on the evaluation set:

$$\text{HTER}_{\tau_*} = \frac{\text{FAR}_{\tau_*} + \text{FRR}_{\tau_*}}{2},$$

where FAR corresponds to the false acceptance rate and FRR the false rejection rate.

As presented in the following section, we adopt HTER as the evaluation measure for both ASVspoof and AVspoof databases.

C. Methodology

We study the proposed approach along with other approaches proposed in the literature in the following manner:

1) we first conduct experiments on the ASVspoof database using the evaluation measure employed in the Interspeech 2015 competition, i.e., EER. We then extend the experiments with HTER as the evaluation measure;
2) next, we conduct experiments on the AVspoof database and study both logical access and physical access attacks with HTER as the evaluation measure;
3) and finally, we investigate the generalization of the systems through cross-database experiments. More specifically, we use the training and development sets of one database to train the system and determine the decision threshold, and then evaluate the systems on the evaluation set of the other database with HTER as the evaluation measure.

D. Systems

In this section, we present the systems investigated, namely, baseline systems and the LTSS based systems. All these

systems have a common preprocessing step for voice activity detection (VAD) to detect the begin and end points of the utterance, which is done by jointly using the normalized log energy and the 4 Hz modulation energy [52] on frame sizes of 20 ms and frame shift of 10 ms. It is worth mentioning that, in case of physical access attacks, this step removes an indicative noise present at the beginning and the end of the utterances, as a consequence of pressing play and stop buttons. Removing those parts ensures that our system is not relying on these portions to differentiate between genuine accesses and attacks.

1) Baseline systems: We selected several state-of-the-art PAD systems that performed well in a recent evaluation by Sahidullah *et al.* [7] on the ASVspoof database as baseline systems.

a) Feature extraction: By following [7], we selected four cepstral-based features with linear-scale triangular (LFCC) and rectangular (RFCC), mel-scale triangular (MFCC) [53], and inverted mel-scale triangular (IMFCC) filters. It is worth pointing out that: (a) RFCC and LFCC only differ in the filter shapes; and (b) LFCC, MFCC, and IMFCC have the same filter shapes but differ in filter placements. These features are computed from a power spectrum (squared magnitude of 512-point fast Fourier transform (FFT)) by applying one of the above filters of a given size (we use size 20 as per [7]). We also implemented spectral flux-based features (SSFC) [52], which are Euclidean distances between power spectrums (normalized by the maximum value) of two consecutive frames, subband centroid frequency (SCFC) [54], and subband centroid magnitude (SCMC) [54] features. A discrete cosine transform (DCT-II) is applied when computing all the above features, except for SCFC, and the first 20 coefficients are taken.

Since Sahidullah *et al.* [7] reported that static features degrade performance of PAD systems, we kept only deltas and double-deltas [55] (40 in total) computed for all features.

b) Classifier: We adopted a GMM-based classifier (two models corresponding to genuine access and attack), since it yielded better systems when compared to SVM. We used the same 512 number of mixtures and 10 EM iterations as done in [7]. The score for each utterance in the evaluation set is computed as a difference between the log-likelihoods of the genuine access model and attack model. The score is thresholded to make the final decision.

2) LTSS-based systems:

a) Feature extraction: The underlying idea of the proposed approach is that the attacks could be detected based on spectral statistics. It is well known that when applying Fourier transform there is a trade-off between time and frequency resolution, i.e., the smaller the frame size, the lower the frequency resolution and the larger the frame size, the higher the frequency resolution. So, the frame size affects the estimation of the spectral statistics.

For both logical access attack and physical access attacks, we determined the frame sizes based on cross validation, while using a frame shift of 10 ms. More precisely, we varied the frame size from 16 ms to 512 ms and chose the frame size that yielded the lowest EER on the development set. For the case of logical access attacks, we have found that frame size of 256 ms yields 0% EER on both ASVspoof and AVspoof databases. In the case of physical access attacks on AVspoof database, we found that 32 ms yields the lowest EER, which is 0.02%. A potential reason for this difference could be that the channel information inherent in physical access attacks is spread across frequency bins while in the case of logical access attacks the relevant information may be localized. We dwell in more detail about it later in Section VI-E.

b) Classifier: We investigate two classifiers, namely, a linear classifier based on linear discriminant analysis (LDA) and a non-linear classifier based on multi-layer perceptron (MLP). The input to the classifiers are the spectral statistics estimated at the utterance level as given in Equation (2) and Equation (3), i.e., one input feature vector per utterance.

LDA: the input features are projected onto one dimension with LDA , i.e., by finding the linear projection of the features components that minimizes intra-class variance and maximizes inter-class variance. We then directly use the values as scores.

MLP: we use an MLP with one hidden layer and two output units. The MLP was trained with a cost function based on the cross entropy using the back propagation algorithm and early stopping criteria. We used the Quicknet software [50] to train the MLP. The number of hidden units was determined through a coarse grid search based on the performance on the development set: 100 hidden units for AVspoof-LA and AVspoof-PA and 10000 hidden units for ASVspoof. During testing we threshold the output probability and make the decision.

We had also carried out investigations using GMMs. However, we do not present those studies as the error rates were significantly higher. This is potentially due to a combination of factors: (a) curse of dimensionality and (b) insufficient data for robust parameter estimation, as we obtain only one feature vector per utterance.

V. Results

This section presents the performance of the different systems investigated. We first present the studies on ASVspoof database in Section V-A, followed by the studies on AVspoof database in Section V-B, and finally the cross database studies in Section V-C.

A. Performance on ASVspoof database

Table III presents the results based on the evaluation protocol used in the ASVspoof 2015 competition. The results for known and unknown attacks of the evaluation set are presented separately. We show the results presented in Table 4 of [7] (columns titled as "*[7] EER (%)*") as well as our Bob-based implementation of the same systems (columns titled as "*Bob EER (%)*"). We can observe that both implementations lead to similar results for known attacks, while our Bob-based system shows smaller error rates for unknown attacks.

Table IV presents the results in terms of HTER. It can be observed that the proposed approach yields the lowest HTERs for known attacks scenario when using a LDA classifier and the lowest HTER for unknown attacks scenario when using a MLP. From the results, it seems that the LDA-based approach

TABLE III
EER(%) OF PAD SYSTEMS ON ASVSPOOF WITH RESULTS IN "*[7] EER (%)*" COLUMN TAKEN FROM [7]. EVALUATION SET.

System	[7] EER (%)		Bob EER (%)	
	Known	Unknown	Known	Unknown
SCFC	0.07	8.84	0.10	5.17
RFCC	0.12	1.92	0.12	1.32
LFCC	0.11	1.67	0.13	1.20
MFCC	0.39	3.84	0.46	2.93
IMFCC	0.15	1.86	0.20	1.57
SSFC	0.30	1.96	0.23	1.60
SCMC	0.17	1.71	0.18	1.37
LTSS, LDA	N/A	N/A	**0.03**	2.09
LTSS, MLP	N/A	N/A	0.10	**0.40**

does not generalize well to unknown attacks. However, as we shall see later in Section VI-B, the performance difference is mainly due to the unknown S10 condition. Furthermore, these results can be contrasted with the results in the right column of Table III, i.e., Bob EER, as they share the same implementation except for the evaluation measure. It can be observed that the rank order of the systems based on the HTER and EER are not the same, especially for the case of unknown attacks. Thus, indicating that choosing the threshold based on the EER of each type of attack in the evaluation set might not a true indicator of practical scenario.

TABLE IV
HTER(%) OF PAD SYSTEMS ON ASVSPOOF. EVALUATION SET.

System	Known	Unknown
SCFC	0.20	6.71
RFCC	0.21	2.11
LFCC	0.27	1.77
MFCC	0.84	3.76
IMFCC	0.32	3.19
SSFC	0.35	2.12
SCMC	0.38	1.88
LTSS, LDA	**0.03**	6.36
LTSS, MLP	0.18	**0.60**

B. Performance on AVspoof database

Table V presents the results on the AVspoof database, which contains both logical access (LA) attacks and physical access (PA) attacks. For each attack type, corresponding baseline and LTSS-based systems were trained and evaluated independently.

TABLE V
HTER (%) OF PAD SYSTEMS ON AVSPOOF, SEPARATELY TRAINED FOR THE DETECTION OF PHYSICAL ACCESS (PA) AND LOGICAL ACCESS (LA) ATTACKS. EVALUATION SET.

System	LA	PA
SCFC	**0.00**	5.15
RFCC	0.03	2.70
LFCC	**0.00**	5.00
MFCC	**0.00**	5.34
IMFCC	0.01	3.76
SSFC	0.70	4.17
SCMC	0.01	3.24
LTSS, LDA	0.04	0.18
LTSS, MLP	1.00	**0.14**

We can note that: (i) the LA set of AVspoof is less challenging compared to ASVspoof for all, except for SSFC-based methods and for our MLP-based system, and (ii) presentation attacks are significantly more challenging compared to LA attacks for all the baseline systems. This means that presentation attacks, besides emulating a more realistic scenario, pose a serious threat to the state of the art systems and need to be considered in all future evaluations of anti-spoofing systems. On the other hand, the proposed approach with linear classifier outperforms the baseline systems on both LA attacks and PA attacks. The MLP-based system yields one of the lowest error rate on PA but performs worse on LA.

C. Cross-database testing

This section presents the study on generalization capabilities of the systems. To do so, as mentioned earlier in Section IV-C, we used the training and development sets of one database and the evaluation set of another database. We train the systems on the detection of logical access attacks and observe whether or not it can generalize to the detection of logical access attacks of another database and to the detection of physical access attacks.

Table VI presents the results of the study. We see that there is no system that outperforms the others in all the scenarios. The performance depends on which data was used during the training and during the evaluation. Furthermore, even though our system outperforms the others when training and evaluating on the same dataset, we observe that it does not generalize well to unseen attacks and unseen recording conditions. Furthermore, LDA-based system outperforms MLP-based system on three scenarios, suggesting that the MLP-based system overfits. We analyze the reasons in Section VI-C.

VI. ANALYSIS

In this section, we give further insights into the long-term spectral statistics based approach. We first compare our approach to systems based on raw log-magnitude spectrum. We then analyze the results obtained on the ASVspoof database per type of attack with a focus on the S10 attack as this is the most challenging one. Afterward, we analyze why our system yields one of the highest error rate in Table VI when trained on the AVspoof-LA database and evaluated on the ASVspoof database. Then, we analyze the LDA classifier to understand the information modeled for logical and physical access attacks. Finally, we study the impact of the frames length, which is directly related to the frequency resolution, on the performance of the system.

A. Comparison to magnitude spectrum-based systems

The raw log-magnitude spectrum computed over short time frames of $\approx 20 - 25$ ms has been used as features in several works, classified either with a GMM [7], a SVM [29], a MLP [11], [12] or with deep architectures [27], [28], [29], [30]. In Table VII, we present the available results on the evaluation set of the ASVspoof database with systems using either raw log-magnitude spectrum ("spec"), filter banks applied after computing the raw log-magnitude spectrum ("fbanks") or with a log-scale spectrogram ("spectro"). "spec + MLP"

TABLE VI
CROSS DATABASE EVALUATION ON ASVSPOOF AND AVSPOOF DATABASES OF PAD SYSTEMS IN TERMS OF HTER (%). EVALUATION SET.

System	ASVspoof (Train/Dev)		AVspoof-LA (Train/Dev)	
	AVspoof-LA (Eval)	AVspoof-PA (Eval)	ASVspoof (Eval)	AVspoof-PA (Eval)
SCFC	1.43	**6.48**	19.99	7.56
RFCC	34.93	38.54	25.58	13.20
LFCC	**0.71**	10.58	18.44	8.40
MFCC	1.87	9.82	**10.13**	**5.15**
IMFCC	2.28	46.49	21.80	49.57
SSFC	34.64	41.68	43.50	36.26
SCMC	1.23	12.16	22.99	7.97
LTSS, LDA	43.35	45.62	14.08	36.64
LTSS, MLP	50.00	50.00	46.13	23.01

corresponds to the results presented in [12], where the raw log-magnitude spectra are classified with a one hidden layer MLP. "fbanks + SVM" and "fbanks + DNN" were presented in [29], filter banks outputs are respectively classified with a SVM and with a 2 hidden layers DNN. "fbanks + DNN [27]" corresponds to filter banks fed to a 5-layers DNN to extract features and classification using Mahalanobis distance. "fbanks + {DNN,RNN}" corresponds to the best system obtained in [30], which is a score-level fusion of features learned with a DNN and classified with a LDA and features learned with a RNN and classified with a support vector machine. The system "spectro + {CNN,RNN,CNN+RNN}" was developed in [28] and is a score-level fusion of a CNN, a RNN and a combined CNN and RNN, all trained on the log-scale spectrogram of the speech utterances. "spec + GMM" and "cep + GMM" correspond to our implementation of log-magnitude spectrum classified with a 512 mixtures GMM. "LTSS + SVM", "LTSS + LR", "LTSS + LDA" and "LTSS + MLP" correspond to long-term spectral statistics based systems with different classifiers: SVM, logistic regression (LR), LDA and MLP, respectively. We can observe that the LTSS linearly classified with LR or LDA outperforms all other systems, even the ones using ANNs with deep architectures to model magnitude spectrum. This shows that indeed the statistics are more informative than the conventional short-term raw log magnitude spectrum, as hypothesized in Section III-A.

Yet, another way to understand these results is through the current trends in ASV, where the state-of-the-art systems are built on top of statistics of cepstral features. More precisely, a GMM-UBM trained with cepstral features is adapted on the speaker data. The parameters, more precisely the mean vectors, of the adapted GMM are then further processed to extract i-vectors, to build systems that are better than the standard GMM-UBM likelihood ratio based system [56]. In our case, we observe a similar trend, i.e., modeling statistics of the raw log magnitude spectrum yields a better system than modeling the raw log magnitude spectrum.

B. Analysis of ASVspoof results

As explained in Section IV-A, the evaluation set of the ASVspoof database contains 10 different types of attacks, denoted respectively S1 to S10, which are either voice conversion or speech synthesis attacks. The attacks S1 to S5 are present in the training, development and evaluation set, while the unknown attacks S6 to S10 are in the evaluation set only.

TABLE VII
EER(%) OF MAGNITUDE SPECTRUM-BASED SYSTEMS ON ASVSPOOF DATABASE. EVALUATION SET.

System	Known	Unknown	Average
spec + MLP [12]	0.06	8.33	4.20
spec + SVM [29]	0.13	9.58	4.85
spec + DNN [29]	0.05	8.70	4.38
fbanks + DNN [27]	0.05	4.52	2.28
fbanks + {DNN,RNN} [30]	0.0	2.2	1.1
spectro + {CNN,RNN,CNN+RNN} [28]	0.27	2.66	1.47
spec + GMM	0.16	3.05	1.60
cep + GMM	0.05	6.23	3.14
LTSS + SVM	0.25	2.70	1.47
LTSS + LR	0.02	1.58	0.80
LTSS + LDA	0.03	2.09	1.06
LTSS + MLP	0.10	0.40	0.25

The attacks S1 to S4 and S6 to S9 are all based on the same "STRAIGHT" vocoder [59], while S5 is based on the MLSA vocoder [60] and S10 is a unit-selection based attack, which does not require any vocoder.

Table VIII shows the per-attack based comparison between the best systems of the Interspeech 2015 ASVspoof competition for which the per-attack results were published (systems "A", "B", "D" and "E"), the best baseline system (LFCC), the recent system based on constant Q cepstral coefficients (CQCC) [58], and the systems based on the proposed LTSS approach. We can observe that all systems achieve very low EERs on the attacks S1 to S9. The main source of error is the S10 attack and the overall performance of the systems differ as a consequence of that. More precisely, among the systems compared, System B and System D in the ASVspoof 2015 challenge yield the best performance across all the attacks except for S10. Similarly, we can see that, in our approach, the LDA classifier consistently yields a comparable or better system than the MLP classifier, except for the S10 attack. This indicates that a more sophisticated classifier is needed to detect attacks arising from concatenative speech synthesis systems. Otherwise, a linear classifier is sufficient to discriminate genuine accesses and attacks based on LTSS. These observations also help in understanding the trends on AVspoof-LA where the LDA based system outperforms the MLP based system.

Finally, it is worth pointing out that in the literature, to the best of our knowledge, CQCC-based approach has achieved the best performance on S10 attack, and as a consequence one of the best overall average performance. We can observe that the proposed LTSS based approach with MLP as classifier

TABLE VIII
EER (%) PER TYPE OF ATTACK COMPUTED ON THE ASVSPOOF DATABASE. EVALUATION SET.

System	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A [20]	0.10	0.86	0.00	0.00	1.08	0.85	0.24	0.14	0.35	8.49
B [57]	0.00	0.02	0.00	0.00	0.01	0.02	0.00	0.02	0.00	19.57
D [11]	0.0	0.0	0.0	0.0	0.01	0.01	0.0	0.0	0.0	26.1
E [21]	0.024	0.105	0.025	0.017	0.033	0.093	0.011	0.236	0.000	26.393
LFCC	0.032	0.500	0.000	0.000	0.126	0.151	0.011	0.234	0.032	5.561
CQCC [58]	0.005	0.106	0.000	0.000	0.130	0.098	0.064	1.033	0.053	1.065
LTSS, LDA	0.000	0.043	0.000	0.000	0.086	0.086	0.022	0.086	0.032	10.218
LTSS, MLP	0.011	0.151	0.000	0.000	0.352	0.288	0.054	0.043	0.065	1.564

closely matches that.

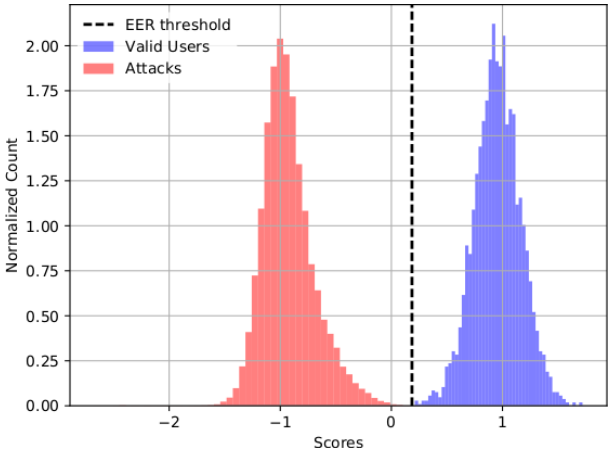
C. Analysis of cross-database performance

In our experimental studies on ASVspoof database, we observed that the proposed approach generalizes across unseen attacks. However, in the case of cross database studies, especially when trained on ASVspoof and tested on AVspoof-LA (see Table VI), we observe that it is worse than all systems. In order to understand, we analyzed the score histograms on ASVspoof that are used to determine the threshold and the score histograms that are obtained in the test condition. Fig. 2 shows these histograms. We see that on the development set of the ASVspoof database, the attacks scores are clearly separated from the genuine accesses scores. However, when applying the same threshold on the evaluation of the AVspoof database, we see that a lot of genuine accesses are wrongly classified as attacks, i.e., the FRR is high (86.496%) while the FAR is still very low (0.002%). We believe that this difference is a consequence of the difference in the recording conditions. Specifically, the genuine speech in ASVspoof database was recorded in a hemi-anechoic chamber using an omni-directional head-mounted microphone. On the other hand, the genuine speech of AVspoof-LA database was recorded in realistic conditions with different microphones: a very good quality microphone, laptop microphone and two smartphones microphones.

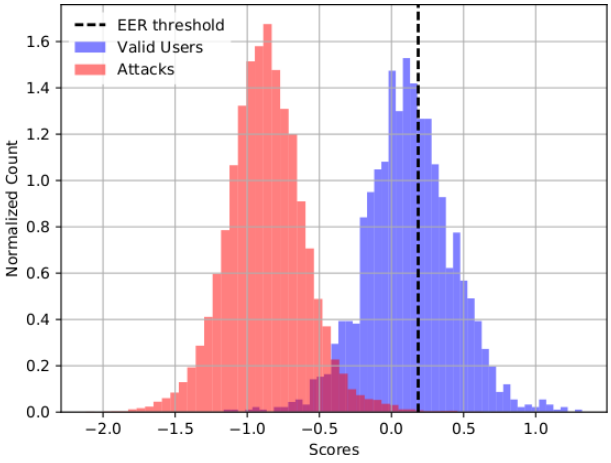
D. Analysis of the discrimination

When classifying the features with a LDA, we project them into one dimension, which best separates the genuine accesses from the attacks in the sense that we maximize the ratio of the "between class variance" to the "within-class variance". By analyzing this projection, we can gain insight about the importance of each component in the original space. More precisely, each extracted feature vector is a concatenation of a spectral mean and a spectral standard deviation. Thus, each half of a feature vector lies in the frequency domain, and their components are linearly spaced between 0 and 8 kHz. For example, if we compute the spectral statistics over frames of 256 ms, each spectral mean and spectral standard deviation vectors are composed of 2048 components and the i^{th} component will correspond to the frequency $\approx i \times 3.91$Hz. Analyzing the LDA projection vector can thus lead us to understand the importance of each frequency region.

Fig. 3 shows the plot of the absolute values of the first 800 components of the projection vector learned by the LDA



(a) Development set: ASVspoof database

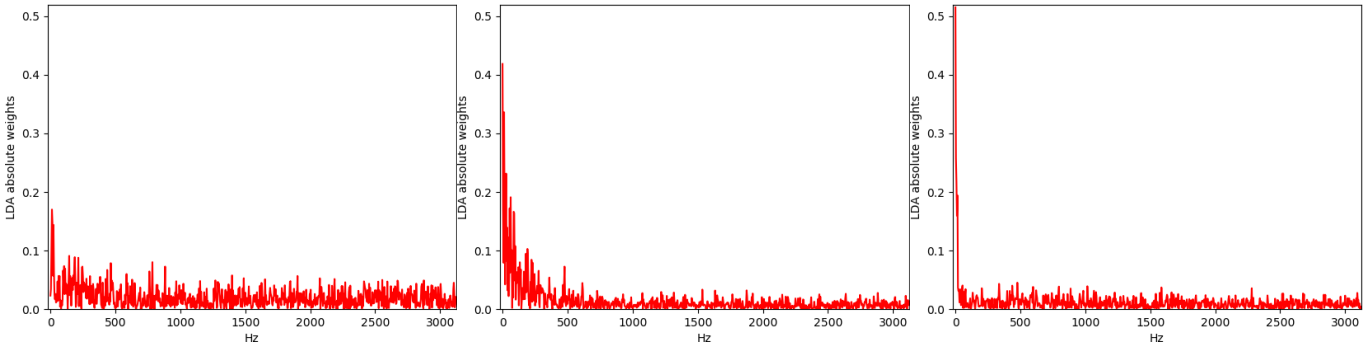


(b) Evaluation set: AVspoof LA database

Fig. 2. *Score histograms of the proposed LDA-based system, trained on the ASVspoof database and evaluated on the AVspoof-LA dataset.*

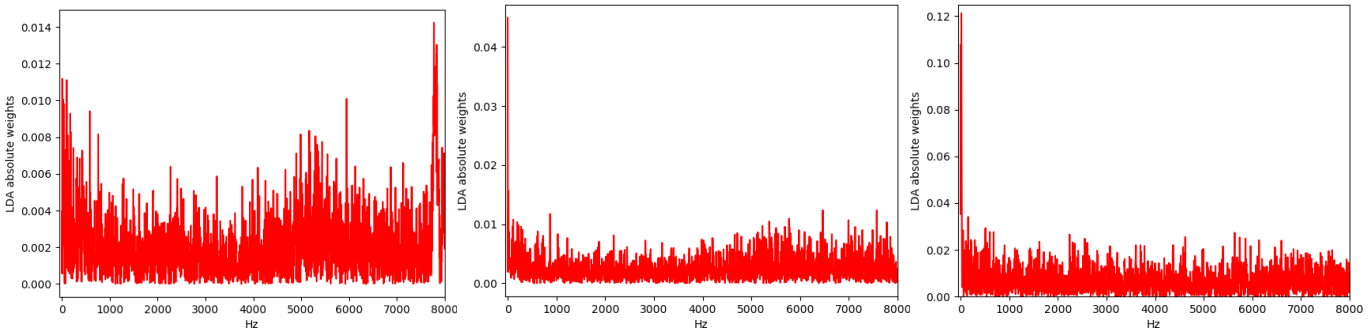
classifier trained to detect the physical access (AVspoof-PA) and logical access (AVspoof-LA) attacks on the AVspoof database, and the logical access attacks on the ASVspoof database (ASVspoof). These components correspond to the spectral mean between 0 and $\approx$ 3128 Hz. As the frequency increase above this value, the average amplitude of the LDA weights does not change, which is why the high-frequency components are not shown on this figure.

We observe that when detecting physical access attacks, even though the weights are slightly higher in the low frequencies, importance is given to all the frequency bins. This can be explained by the fact that playing the fake sample through



(a) physical access attacks (AVspoof PA) (b) logical access attacks (AVspoof LA) (c) logical access attacks (ASVspoof)

Fig. 3. *800 first LDA weights for physical and logical attacks of AVspoof and ASVspoof databases, corresponding to the frequency range* $[0, 3128]$ *Hz of the spectral mean.*



(a) physical access attacks (AVspoof-PA) (b) logical access attacks (AVspoof-LA) (c) logical access attacks (ASVspoof)

Fig. 4. *LDA weights corresponding to the spectral standard deviation for physical and logical attacks of AVspoof and ASVspoof databases.*

loudspeakers will modify the channel impulse response across the whole bandwidth. Thus, the relevant information to detect such attacks is spread across all frequency bins. However, in the case of logical access attacks, we observe that the largest weights correspond to a few frequency bins that are well below 50 Hz, i.e., the discriminative information in the frequency domain is highly localized in the low frequencies.

Figure 4 presents the LDA weights corresponding to the spectral standard deviation. The observations are similar to the ones made on the spectral mean. For the detection of physical access attacks, i.e., on AVspoof-PA, the information is spread across all the frequencies. On the other hand, in the case of logical access attacks, i.e., on AVspoof-LA and ASVspoof, the emphasis is given to the low frequencies. Furthermore we can observe that the LDA weights are relatively smaller when compared to the spectral mean. This suggests that the mean is more discriminative than the standard deviation. To confirm this hypothesis, we conducted an investigation using stand-alone features. Table IX presents the results. It can be seen that the stand-alone mean (μ) features yields a better system than the stand-alone standard deviation (σ) features, including in cross-database scenarios (systems trained on ASVspoof and evaluated on AVspoof-LA and conversely). The combined feature leads to a better system, except on ASVspoof known attacks.

One explanation for the importance of the low frequency region for the detection of logical access attacks could be the following. Natural speech is primarily realized by movement of articulators that convert DC pressure variations created during respiration into AC pressure variations or speech sounds [61]. Alternatively, there is an interaction between pulmonic and oral systems during speech production. In speech processing, including speech synthesis and voice conversion, the focus is primarily on glottal and oral cavity through source-system modeling. In the proposed LTSS-based approach, however, no such assumptions are being made. As a consequence, the proposed approach could be detecting logical access attacks on the basis of the effect of interaction between pulmonic and oral systems that exists in the natural speech but not in the synthetic or voice converted speech (due to source-system modeling and subsequent processing). It is understood that the interaction between pulmonary and oral cavity systems can create DC effects when producing sounds such as clicks, ejectives, implosives [61]. Furthermore, human breath in the respiration process can reach the microphone and appear as "pop noise" [26], which again manifests in the very low

TABLE IX
IMPACT OF THE MEAN AND STANDARD DEVIATION FEATURES USED STAND-ALONE AND COMBINED.

	AVspoof PA	**AVspoof LA**	**ASVspoof known / unknown**	**ASVspoof (Train) AVspoofLA (Eval)**	**AVspoofLA (Train) ASVspoof (Eval)**
μ	0.51	0.04	0.02 / 6.96	45.56	26.25
σ	2.03	4.65	4.10 / 19.46	55.42	45.15
$[\mu, \sigma]$	0.18	0.04	0.03 / 6.36	43.35	14.08

frequency region. Finally, it is worth mentioning that our observations are somewhat different than the observations made in [62], [63], where the authors have observed that high frequency regions were also helping in discriminating natural speech against synthetic speech. This difference can be due to the manner in which the signal is modeled and analyzed. In [62], [63], the analysis has been carried out with standard short-term speech processing, while in our case the analysis is carried out on statistics of log magnitude spectrum of 256 ms signal. So the importance of high frequency in standard short-term speech processing could be due to the differences in the spectral characteristics of specific speech sounds (e.g. fricatives) in genuine speech and synthetic speech. In our case, the speech sound information is averaged out.

E. Analysis of the impact of the frame length

In the experimental studies, we observed that physical access attacks and logical access attacks need two different window sizes (found through cross-validation). A question that arises is: what is the role of window size or frame lengths in the proposed approach? In order to understand that, we performed evaluation studies by varying the frame lengths: 16ms, 32 ms, 64 ms, 128 ms, 256 ms and 512 ms with a frame shift of 10 ms. The length of each feature is $2^{\lceil log_2 w_l \rceil}$. For example, a frame length of 32 ms will yield features of 512 components. Fig. 5 presents the HTER computed on the evaluation set for different frame lengths. We compare the performance impact on the detection of physical and logical access attacks of the AVspoof database and on the logical access attacks of the ASVspoof database. For the sake of clarity, unknown S10 attack results are presented separately than the rest if unknown attacks S6-S9.

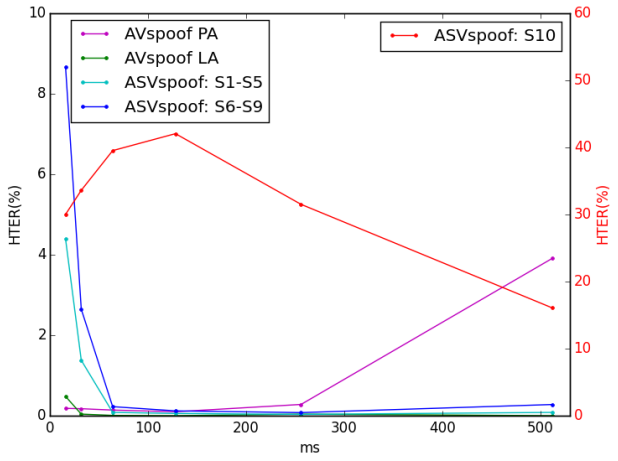


Fig. 5. *Impact of frames lengths on the performance of the proposed LDA-based approach, evaluated on the three datasets: ASVspoof, AVspoof-LA and AVspoof-PA.*

For physical access attacks AVspoof-PA, it can be observed that the HTER slightly decreases from 16 ms to 128 ms and after that it degrades. A likely reason for the degradation after 128 ms is that in physical access attacks there is a channel effect. For that effect to be separable and meaningful for the task at hand, the channel needs to be stationary. We speculate that the stationary assumption is not holding well on longer window sizes.

For logical access attacks, it can be observed that for AVspoof-LA, ASVspoof S1-S5 (known) and ASVspoof S6-S9 (unknown), the HTER steadily drops from 16 ms until 256 ms with a slight increase at 512 ms. Whilst for ASVspoof S10, which contains attacks synthesized using unit selection speech synthesis system, the performance degrades at first and then steadily drops with increase of window size. This could be due to the fact that long-term temporal information is important to detect concatenated speech, since artefacts can happen at the phoneme joint areas. Our results indicate that for attacks based on parametric modeling of speech, as in the case of ASVspoof S1-S9 and AVspoof-LA, frequency resolution is not an important factor while for unit selection based concatenative synthesis, where the speech is synthesized by concatenating speech waveforms, high frequency resolution is advantageous or helpful. More specifically, together with the observations made in the previous section, we conclude that the relevant information to discriminate genuine access and logical access attacks based on concatenative speech synthesis is highly localized in the low frequency region. This conclusion is in line with the observations made with the use of CQCC feature [58], which also provides high frequency resolution in the low frequency regions and leads to large gains on S10 attack condition.

Building on these observations, we asked a question: what is the impact of window length on modeling raw log-magnitude spectrum features? We conducted an experiment, where similar to the analysis, the window length was varied as 16ms, 32ms, 64ms, 128ms and 256ms, always shifted by 10ms, and the raw log-magnitude spectrum was modeled by 512-components GMM. The EERs obtained on the evaluation set of the ASVspoof database are shown in Table X. We can observe that for statistical parametric speech synthesis based attacks (S1-S9), the optimal frame length is 64ms, while it is 128ms for unit-selection based attacks (S10). Hypothetically, increase of window size should converge towards LTAS, as it would average out phonetic structure information. However, when compared to spectral statistics, increasing the window size beyond 128 ms starts degrading the performance. This could be potentially due to the difficulty in modeling discriminative information in the high dimensional raw log magnitude spectrum. In fact, in the present study modeling raw log magnitude spectrum of 512ms window became prohibitive both in terms of storage and computation.

Taken together, these analyses clearly show that typical short-term speech processing with 20-30 ms window size and other speech signal related assumptions such as source-system modeling is not a must for detecting presentation attacks.

VII. CONCLUSIONS

In one of the first efforts, this paper investigated in depth the detection of both physical access attacks and logical access attacks. In this context, we proposed a novel approach

TABLE X
IMPACT OF FRAMES LENGTHS ON THE PERFORMANCE OF RAY LOG-MAGNITUDE SPECTRUM CLASSIFIED WITH A GMM. EER(%) OF EVALUATION SET OF ASVSPOOF DATABASE.

frames length (ms)	Known	Unknown		Average
		S6-S9	S10	
16	0.11	0.10	17.95	1.89
32	0.06	0.08	18.59	1.92
64	0.04	0.04	8.97	0.94
128	0.05	0.05	6.81	0.72
256	0.06	0.09	8.26	0.89

that detects presentation attacks based on the input signal magnitude spectrum statistics and studied it in comparison to approaches based on conventional short-term spectral features. Our investigations on two separate datasets, namely, ASVspoof 2015 challenge and AVspoof led to the following observations:

1) The proposed approach, which does not make any speech signal modeling related assumptions, works equally well for both physical access attacks and logical access attacks. However, analysis of the linear discriminative classifier shows that for physical access attacks the discriminative information is spread over different frequency bins while for logical access attacks the discriminative information is more localized in low frequency bins.
2) Standard short-term spectral features based approaches proposed in the literature work well for logical access attacks but lead to inferior systems on physical access attacks, when compared to the proposed approach. This can be due to the fact that in the literature the research has mainly focused on logical access attacks. As a consequence, the methods may be more tuned to that.
3) Cross-database and cross-attack studies suggest that long-term spectral statistics based approach do not generalize well. The cross-attack aspect is understandable given that the classifier models different information for physical access attacks and logical access attacks. Our studies also show that none of the approaches based on standard short-term spectral processing truly generalize across databases. Such a claim arises as, despite observing that the LFCC-based system trained on ASVspoof leads to a low HTER on AVspoof-LA test set, a small modification, i.e., by just replacing the triangular shaped filters by rectangular shaped ones leads to a drastic degradation.

Taken together these observations provide the following directions for future research:

1) The proposed approach of using long-term spectral statistics provides benefits such as a simple feature extraction with no speech signal related assumption and a linear classifier. So, should we treat the problem from the perspective of prior knowledge based speech processing or not? In that direction, we aim to focus on up-and-coming deep learning based approaches that in a data- and task-driven manner determines the appropriate block processing and learns the relevant features and the classifier jointly from the raw speech signal [64], [65]. Such methods of discovering features could lead to better understanding of the problem.
2) The cross-domain studies show that there is a need for more resources and further research on how to make the counter-measure systems robust or domain invariant. Furthermore, LTSS-based approach has been investigated in relatively clean conditions. Further investigations are needed to ascertain their benefit in adverse conditions. In our future work, we aim to explore multiple classifier fusion techniques in these directions, as the studies indicate that a single feature would not be sufficient.
3) Our experiments and analyses show that physical access attacks and logical access attacks are not of the same nature. So should the future research emphasis lie on physical access attacks or logical access attacks? Given the realistic nature of physical access attacks, our future work will build on the on-going initiatives in the context of the SWAN project[6] for data collection and development of counter-measures as well as in the context of ASVspoof 2017[7] challenge.

ACKNOWLEDGMENT

This project was partially funded by the Swiss National Science Foundation (project UniTS), the Research Council of Norway (project SWAN) and the European Commission (H2020 project TESLA).

REFERENCES

[1] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, Mar. 2001.
[2] S. Kucur Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2015.
[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.
[4] ISO/IEC JTC 1/SC 37 Biometrics, "DIS 30107-1, information technology – biometrics presentation attack detection," American National Standards Institute, Jan. 2016.
[5] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?" Idiap Research Institute, Tech. Rep. Idiap-RR-61-2005, 2005.
[6] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. of Interspeech*, 2015.
[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. of Interspeech*, 2015.
[8] A. Anjos, L. El-Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *Proc. of the ACM International Conference on Multimedia*, 2012.
[9] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *Proc. of International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2016.
[10] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Proc. of Interspeech*, 2016.

[6]https://www.ntnu.edu/aimt/swan
[7]http://www.spoofingchallenge.org/

[11] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for asvspoof 2015 challenge," in *Proc. of Interspeech*, 2015.

[12] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[13] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[14] P. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[15] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition." in *Proc. of Interspeech*, 2012.

[16] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. of ICASSP*. IEEE, 2013, pp. 7234–7238.

[17] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. of BTAS*, Sept 2013, pp. 1–8.

[18] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. of Interspeech*, 2013.

[19] J. Gaka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[20] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. of Interspeech*, 2015.

[21] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. of Interspeech*, 2015.

[22] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. of Interspeech*, 2015.

[23] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," *Proc. of Interspeech*, 2015.

[24] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis." in *Proc. of Interspeech*, 2012.

[25] A. Ogihara, U. Hitoshi, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 1, pp. 280–286, 2005.

[26] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *Proc. of Interspeech*, 2015.

[27] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection - the SJTU system for ASVspoof 2015 challenge," in *Proc. of Interspeech*, 2015.

[28] C. Zhang, C. Yu, and J. H. Hansen, "An investigation of deep learning frameworks for speaker verification anti-spoofing," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[29] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. of Interspeech*, 2015.

[30] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, pp. 43–52, 2016.

[31] T. Kinnunen, V. Hautamäki, and P. Fränti, "On the use of long-term average spectrum in automatic speaker recognition," in *Proc. of Int. Symposium on Chinese Spoken Language Processing*. Citeseer, 2006.

[32] A. Löfqvist, "The long-time-average spectrum as a tool in voice research," *Journal of Phonetics*, vol. 14, pp. 471–475, 1986.

[33] H. K. Dunn and S. D. White, "Statistical measurements on conversational speech," *Journal of the Acoustical Society of America*, vol. 11, pp. 278–288, January 1940.

[34] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[35] K. Tanner, N. Roy, A. Ash, and E. H. Buder, "Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy?" *Journal of Voice*, vol. 19, no. 2, pp. 211–222, 2005.

[36] L. K. Smith and A. M. Goberman, "Long-time average spectrum in individuals with parkinson disease," *NeuroRehabilitation*, vol. 35, no. 1, pp. 77–88, 2014.

[37] S. Master, N. d. Biase, V. Pedrosa, and B. M. Chiari, "The long-term average spectrum in research and in the clinical practice of speech therapists," *Pró-Fono Revista de Atualização Científica*, vol. 18, no. 1, pp. 111–120, 2006.

[38] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS)," *Journal of Voice*, vol. 10, no. 1, pp. 59–66, 1997.

[39] S. E. Linville and J. Rens, "Vocal tract resonance analysis of aging voice using long-term average spectra," *Journal of Voice*, vol. 15, no. 3, pp. 323–330, 2001.

[40] T. Leino, "Long-term average spectrum study on speaking voice quality in male actors," in *Proc. of the Stockholm Music Acoustics Conference*, 1993.

[41] J. Sundberg, "Perception of singing," *The psychology of music*, vol. 1999, pp. 171–214, 1999.

[42] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Proceedings of Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 113–118.

[43] T. Drugman and T. Raitio, "Excitation modeling for HMM-based speech synthesis: Breaking down the impact of periodic and aperiodic components," in *Proceedings of ICASSP*, 2014.

[44] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking," in *Proc. Symp. on Time Series Analysis*, 1963.

[45] A. V. Oppenheim and R. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[46] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[47] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1217–1232, 2012.

[48] M. H. Soni and H. A. Patil, "Non-intrusive quality assessment of synthesized speech using spectral features and support vector regression," in *Proceedings of 9th ISCA Speech Synthesis Workshop*, 2016, pp. 139–145.

[49] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001, pp. 517–519.

[50] D. Johnson *et al.*, "ICSI Quicknet Software Package," http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[51] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. G. S. Mello, R. P. V. Violato, F. O. Simões, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul, G. Saha, and M. Sahidullah, "Overview of BTAS 2016 speaker anti-spoofing competition," in *BTAS*, 2016.

[52] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of ICASSP*, 1997.

[53] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[54] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. C. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, Apr. 2011.

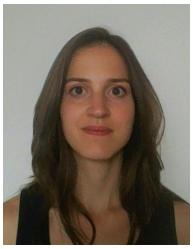
[55] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[56] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[57] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *Proc. of ICASSP*, 2016.

[58] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. of Odyssey*, 2016.

[59] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.
[60] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. of ICASSP*, 1992.
[61] J. J. Ohala, "Respiratory activity in speech," in *Speech Production and Speech Modeling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990.
[62] D. Paul, M. Pal, and G. Saha, "Spectral features for synthetic speech detection," *IEEE Journal of Selected Topics in Signal Processing*, 2017.
[63] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech," in *Proc. of Interspeech*, 2016.
[64] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks," in *Proc. of Interspeech*, September 2013.
[65] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition," Idiap Research Institute, Tech. Rep. Idiap-RR-18-2016, 2016, submitted to Speech Communication.



Hannah Muckenhirn (S'17) received the Master of Science (M.S.) in Communication Systems from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2014. She worked as a data scientist in the industry, focusing mainly on natural language processing and smart meter data analytics. She is currently a research assistant at the Idiap Research Institute, Martigny, Switzerland and pursuing a PhD in Electrical Engineering at EPFL, Switzerland. Her thesis research is focusing on building trustworthy speaker recognition systems. Her research interest includes machine learning, speech and audio signal processing.

Pavel Korshunov received his Ph.D. in Computer Science from National University of Singapore in 2011 and was a postdoctoral researcher at EPFL, Switzerland (2011–2015). He is now a research associate in Biometrics group at the Idiap Research Institute (CH), working on speaker presentation attack detection and detection of inconsistencies between audio and video. He is also a contributor to the signal processing and machine learning open source toolbox Bob. He is a recipient of ACM TOMM Nicolas D. Georganas Best Paper Award in 2011, two top 10% best paper awards in MMSP 2014, and top 10% best paper award in ICIP 2014. He has over 50 research publications and is a co-editor of the JPEG XT standard for HDR images. His research interests include computer vision, crowdsourcing, high dynamic range imaging, speech and video analysis, biometrics, privacy protection, tampering detection, and machine learning.

Mathew Magimai.-Doss (S '03, M'05) received the Bachelor of Engineering (B.E.) in Instrumentation and Control Engineering from the University of Madras, India in 1996; the Master of Science (M.S.) by Research in Computer Science and Engineering from the Indian Institute of Technology, Madras, India in 1999; the PreDoctoral diploma and the Docteur s Sciences (Ph.D.) from the Ecole Polytechnique Fdrale de Lausanne (EPFL), Switzerland in 2000 and 2005, respectively. He was a postdoctoral fellow at the International Computer Science Institute (ICSI), Berkeley, USA from April 2006 till March 2007. Since April 2007, he has been working as a Researcher at the Idiap Research Institute, Martigny, Switzerland. He is also a lecturer at EPFL, where he teaches courses on speech and audio processing. He is a Senior Area Editor of the IEEE Signal Processing Letters. He was an Associate Editor of the IEEE Signal Processing Letters (2013-2017). His main research interest lies in signal processing, statistical pattern recognition, artificial neural networks and computational linguistics with applications to speech and audio processing and multimodal signal processing.

Sébastien Marcel received the Ph.D. degree in signal processing from Université de Rennes I in France (2000) at CNET, the research center of France Telecom (now Orange Labs). He is currently interested in pattern recognition and machine learning with a focus on biometrics security. He is a senior researcher at the Idiap Research Institute (CH), where he heads a research team and conducts research on face recognition, speaker recognition, vein recognition and presentation attack detection (anti-spoofing). He is lecturer at the Ecole Polytechnique Fédérale de Lausanne (EPFL) where he is teaching on "Fundamentals in Statistical Pattern Recognition". He is Associate Editor of IEEE Signal Processing Letters. He was Associate Editor of IEEE Transactions on Information Forensics and Security, a Co-editor of the "Handbook of Biometric Anti-Spoofing", a Guest Editor of the IEEE Transactions on Information Forensics and Security Special Issue on "Biometric Spoofing and Countermeasures", and Co-editor of the IEEE Signal Processing Magazine Special Issue on "Biometric Security and Privacy". Finally he was the principal investigator of international research projects including MOBIO (EU FP7 Mobile Biometry), TABULA RASA (EU FP7 Trusted Biometrics under Spoofing Attacks) and BEAT (EU FP7 Biometrics Evaluation and Testing).