IDIAP RESEARCH REPORT

LONG TERM SPECTRAL STATISTICS FOR VOICE PRESENTATION ATTACK DETECTION

Hannah Muckenhirn [a] Pavel Korshunov
Mathew Magimai.-Doss Sébastien Marcel

Idiap-RR-11-2017

MARCH 2017

[a] Idiap Research Institute

Long Term Spectral Statistics for Voice Presentation Attack Detection

Hannah Muckenhirn[1,2], Pavel Korshunov[1], Mathew Magimai.-Doss[1], Sébastien Marcel[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

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 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 Interspeech 2015 ASVspoof challenge dataset and AVspoof dataset show that, when compared to the approaches based on conventional short-term spectral processing, 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.

1 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. Though the ultimate goal is to have a system that is error free, the ASV systems in practice 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 authentication system, ASV-based authentication systems, or in general biometric systems, can be attacked. Specifically, as illustrated in Figure 1, there are different points at which a biometric system can be attacked [1]. In this paper, our interest lies in attacks at point (1) and point (2), 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. Prevention of or countering such attacks is more related to cyber-security, and is thus out of the scope of the present paper.

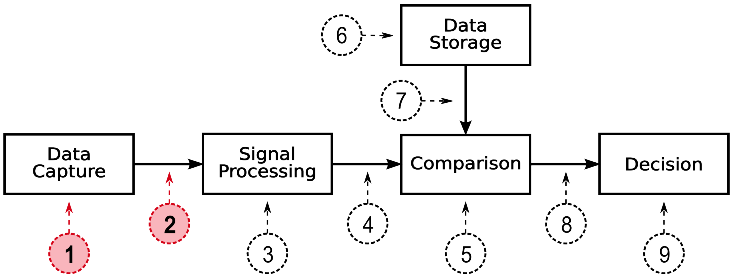


Figure 1: *Potential points of attack in a biometric system, as defined in the ISO-standard 30107-1 [4]. Points 1 and 2 correspond respectively to attacks performed via physical and via logical access.*

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 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 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 undesirable variabilities that ASV systems are robust to and thus, can pass undetected.

As a consequence, developing countermeasures to detect spoofing 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 2, 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. However, both genuine accesses and presentation attacks are speech signals that carry same high level information, such as message, speaker identity, and information about environment. There is not much prior knowledge that can guide us to differentiate between genuine access speech from presentation attack speech. So

a question that arises is: do we still need to follow standard short-term speech processing techniques for PAD? In that direction, we propose a novel approach that simply uses first order and second order spectral statistics computed over Fourier magnitude spectrum to detect presentation attacks.

2. As mentioned earlier, research on detecting spoofing attacks has mainly focussed 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 from it 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][1], 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]. We focus on the analysis of these results and show that the models learned for the detection of logical and physical access attacks are different and that, as a consequence, models cannot generalize.

The remainder of the paper is organized as follows. Section 2 provides a background on the countermeasures developed for logical access attacks. Section 3 then motivates and presents the proposed spectral statistic based approach for PAD. Section 4 presents the experimental setup. Section 5 presents the results and Section 6 presents an analysis of the proposed approach and results obtained. Finally, in Section 7, we conclude.

2 Related work

As mentioned earlier, various methods have been proposed in the context of logical access attack detection. All these approaches, as illustrated in Figure 2, 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].

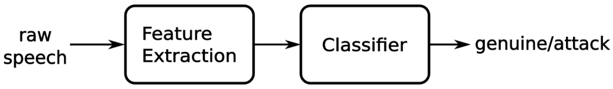


Figure 2: *Presentation attack detection system.*

2.1 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 [7]: this includes standard cepstral features (e.g., mel frequency cepstral coefficients, perceptual linear prediction cepstral coefficients, linear prediction cepstral coefficients), spectral flux-based features

[1] http://idiap.github.io/bob/

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 [11, 12, 7]: this includes group delay-based features, cosine-phase function, and relative phase shift.

3. spectral-temporal features: this includes modulation spectrum [7], frequency domain linear prediction [7], extraction of local binary patterns in the cepstral domain [13, 14], and spectrogram based features [15].

The magnitude spectrum based features and phase spectrum based features have been investigated individually as well as in combination [16, 17, 18, 19].

In addition to these spectral-based features, features based on pitch frequency patterns have been proposed [20, 21]. There are also methods that aim to extract "pop-noise" related information that is indicative of the breathing effect inherent in normal human speech [22].

2.2 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, 13], artificial neural networks (ANNs) [23, 24], and Gaussian mixture models (GMMs) [7, 11, 12, 16, 17, 18, 19]. 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 [25], 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 GMM-UBM ASV system, and is then compared to a preset threshold to make the final decision.

3 Proposed approach: long-term spectral statistics

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

3.1 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 the previous section, 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.

When such measurements are done in the conventional short-term framework, i.e., using 20-30 ms frame sizes, they could be meaningful for detection of presentation attacks. For instance, long term average spectrum (LTAS) has been used in the clinical domain for voice quality measurement. It is employed for example for the early detection of voice pathology [26] or Parkinson disease [27], or for evaluating the effect of speech therapy or surgery on the voice quality [28]. In addition to assessing voice quality, LTAS has also been used to investigate voice characteristics. For example, to differentiate between speakers gender [29] and speakers age [30], and also to study singers and actors voices [31, 32]. Natural speech and synthetic speech differ in voice quality. This difference could be captured in terms of spectral statistics.

The long-term spectral statistics are also used to build robust speech and speaker recognition systems. Specifically, state-of-the-art speech and speaker recognition systems employ cepstral mean normalization (CMN) [33] and cepstral variance normalization (CVN) [34] to handle channel variability. Formally, the cepstrum is the Fourier transform of the log magnitude spectrum [35, 36]. Thus, mean and variance of log magnitude spectrum is indicative of a channel variability, which is a desirable feature for presentation attack detection.

In summary, as spectral statistics can be indicative of voice quality as well as channel variability, we hypothesize that they can be used to develop countermeasures against presentation attacks. The following section presents the approach in detail.

3.2 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. In our approach, it is a hyper parameter that is determined through cross validation.

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:

$$\mu[k] = \frac{1}{M} \sum_{m=1}^{M} \log |X_m[k]|, \tag{2}$$

$$\sigma^2[k] = \frac{1}{M} \sum_{m=1}^{M} (\log |X_m[k]| - \mu[k])^2, \tag{3}$$

$k = 0 \cdots \frac{N}{2} - 1$.

This step yields a single vector representation for each input signal or 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 access 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.

4 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[3] and are reproducible[4].

4.1 Databases

We present experiments on two databases, which are the largest speech databases that can be used for attack detection: (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.

4.1.1 ASVspoof

The ASVspoof[5] 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.

4.1.2 AVspoof

The AVspoof database[6] contains replay attacks, as well as speech synthesis and voice conversion attacks both produced via logical and physical access. This database contains the recording 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 a laptop used for the automatic speaker verification 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.

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

[3] http://www1.icsi.berkeley.edu/Speech/qn.html

[4] Source code: https://pypi.python.org/pypi/bob.paper.taslp_2016

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

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

4.2 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 1 and in Table 2 respectively for the ASVspoof database and the AVspoof database.

Table 1: 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 2: 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

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

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

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.

EER is an unrealistic evaluation measure as the performance is measured based on a decision threshold determined on the evaluation set. Thus, 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:

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

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

4.3 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.

4.4 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 [37] 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.

4.4.1 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.

Feature extraction By following [7], we selected four cepstral-based features with linear-scale triangular (LFCC) and rectangular (RFCC), mel-scale triangular (MFCC) [38], 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 (power of magnitude of 512-sized 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) [37], which are Euclidean distances between power spectrums (normalized by the maximum value) of two consecutive frames, subband centroid frequency (SCFC) [39], and subband centroid magnitude (SCMC) [39] features. A discrete cosine transform (DCT-II) is applied to 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 [40] (40 in total) computed for all features.

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 finally thresholded to make the final decision.

4.4.2 LTSS-based systems

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. Alternately, the frame size is an hyper-parameter.

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 yielded 0% EER on both ASVspoof Challenge and AVspoof database. 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 6.4.

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 and we 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 mean square error using the back propagation algorithm and early stopping criteria. The number of hidden units was determined based on cross validation.

5 Results

This section presents the performance of the different systems investigated. We first present the studies on ASVspoof database in Section 5.1 followed by the studies on AVspoof database in Section 5.2, and finally the cross database studies in Section 5.3.

5.1 Performance on ASVspoof database

For the purpose of reference, Table 3 shows the five best systems (denoted as System ID A-E) proposed in the ASVspoof 2015 challenge [6]. The ASVspoof 2015 challenge systems typically employed multiple features and fusion techniques. For example, the team that achieved the best performance [16] used a fusion of cochlear filter cepstral coefficients, instantaneous frequency and mel-frequency cepstral coefficients, classified with a GMM. Similarly, the second best system [41] employed fusion of multiple features based on mel-frequency cepstrum and phase spectrum; transforming them into i-vectors; and finally classifying with a support vector machine. More information on these systems can be found in the respective citations provided in the table.

Table 3: Performance of the five best systems in the ASVspoof 2015 challenge in terms of EER (%), taken from [6]. Evaluation set.

System	Known	Unknown	Average
A [16]	0.408	**2.013**	1.211
B [41]	0.008	3.922	1.965
C [23]	0.058	4.998	2.528
D [24]	**0.003**	5.231	2.617
E [17]	0.041	5.347	2.694

Table 4 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. Furthermore, when compared to the top five systems in the ASVspoof 2015 challenge (Table 3), it can be observed that the baseline systems typically yield lower performance on known attacks but largely better performance on unknown attacks. The proposed approach however yields comparable systems on known attacks and comparable or better systems on unknown attacks.

Table 4: 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**

Table 5 presents the results in terms of HTER. It can be observed that the proposed approach yields one of the lowest HTERs for known attacks scenario when using a LDA classifier and the lowest HTER for unknown attacks scenario when using a MLP. Furthermore, these results can be contrasted with the results in the right column of Table 4, 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 same, especially for the case of unknown attacks. Thus, indicating that the evaluation based on EER is not a true indicator of practical scenario.

Table 5: 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**

5.2 Performance on AVspoof database

Table 6 presents the results on the AVspoof database, which contains both logical access (LA) attacks and physical access (PA) attacks. Two separate baseline systems and LTSS-based systems were trained and evaluated for these two attacks.

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. The increase in HTER for almost all the baseline PAD systems is considerably high for PA compared to LA attacks. 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.

Table 6: 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**

Table 7: 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

5.3 Cross-database testing

This section presents the study on generalization capabilities of the systems. To do so, as mentioned earlier in Section 4.3, we used the training and development sets of one database and the evaluation set of anther 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 7 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 6.2.

6 Analysis

In this section, we give further insights into the long-term spectral statistics based approach. We first 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. Then, we analyze why our system yields one of the highest error rate in Table 7 when trained on the AVspoof-LA database and evaluated on the ASVspoof database. Afterwards, 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.

Table 8: EER (%) per type of attack computed on the ASVspoof database. Evaluation set.

System	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A [16]	0.10	0.86	0.00	0.00	1.08	0.85	0.24	0.14	0.35	8.49
B [41]	0.00	0.02	0.00	0.00	0.01	0.02	0.00	0.02	0.00	19.57
D [24]	0.0	0.0	0.0	0.0	0.01	0.01	0.0	0.0	0.0	26.1
E [17]	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 [42]	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

6.1 Analysis of ASVspoof results

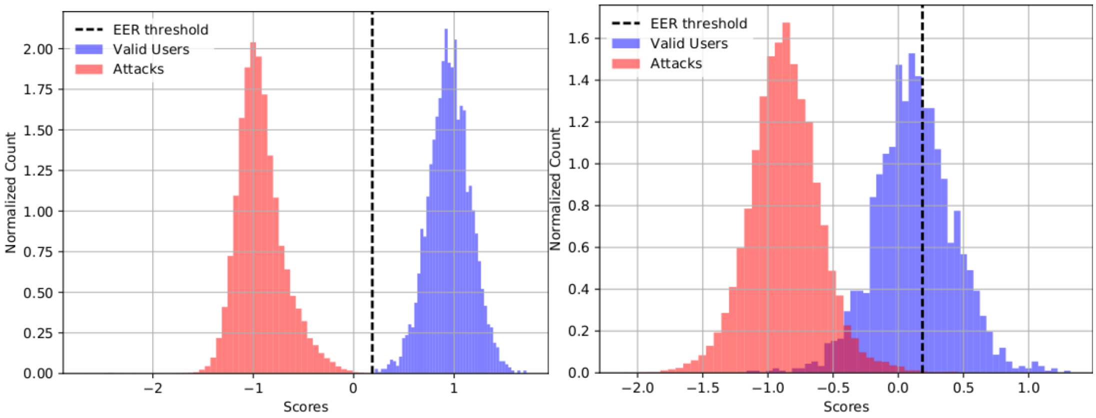
As explained in Section 4.1, the evaluation set of the ASVspoof database contains 10 different types of attacks, denoted respectively S1 to S10, which are either voice conversion of speech synthesis attacks. The attacks S1 to S5 are contained in the training, development and evaluation set, while the attacks S6 to S10 are only in the evaluation set. The attacks S1 to S4 and S6 to S9 are all based on the same "STRAIGHT" vocoder [43]. On the other hand, S5 is based on the MLSA vocoder [44], and S10 is a unit-selection based attack, which does not require any vocoder.

Table 8 shows the per-attack based comparison between the best systems of the Interspeech 2015 ASVspoof competition, the best baseline system (LFCC), the recent system based on constant Q cepstral coefficients (CQCC) [42], 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 D and System B in the ASVspoof Interspeech 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 closely matches that.

6.2 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 7), 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. Figure 3 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.



(a) Development set: ASVspoof database

(b) Evaluation set: AVspoof LA database

Figure 3: *Score histograms of the proposed LDA-based system, trained on the ASVspoof database and evaluated on the AVspoof-LA dataset.*

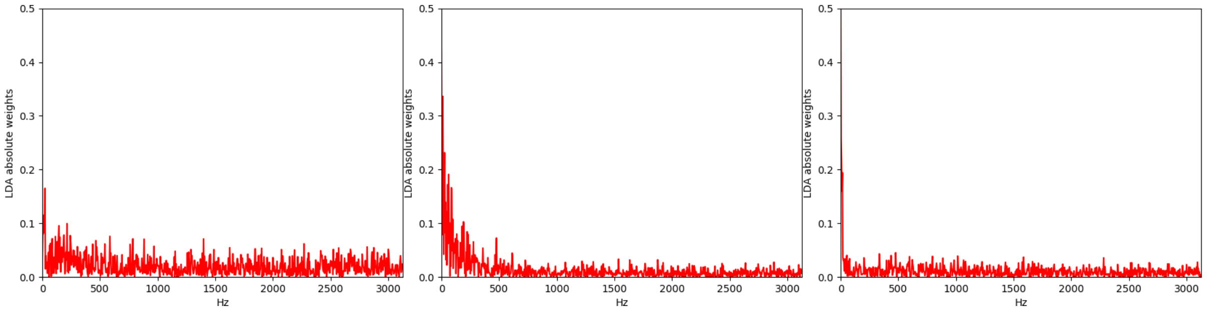
6.3 Analysis of the discrimination

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.

Figure 4 shows the plot of the absolute values of the first 800 components of the projection vector learned by the LDA 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 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 importance is given 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. We observed similar trend as AVspoof-LA with LDA classifier trained on ASVspoof database.

Natural speech is primarily realized by movement of articulators that convert DC pressure variations created during respiration into AC pressure variations or speech sounds [45]. Alter-



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

Figure 4: *800 first LDA weights for physical and logical attacks of AVspoof and ASVspoof databases, corresponding to the frequency range* $[0, 3128]$ *Hz.*

natively, 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 [45]. Furthermore, human breath in the respiration process can reach the microphone and appear as "pop noise" [22], which again manifests in the very low frequency region. This possibly explains why the LDA classifier gives emphasis to very low frequency regions in the case of logical access attacks.

6.4 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, namely, 16ms, 32 ms, 64 ms, 128 ms, 256 ms and 512 ms. Figure 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.

For physical access attacks AVspoof-PA, it can be observed that the performance steadily 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 performance steadily drops from 16 ms till 256 ms with slight degradation 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. Our results indicate that for attacks arising due to 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

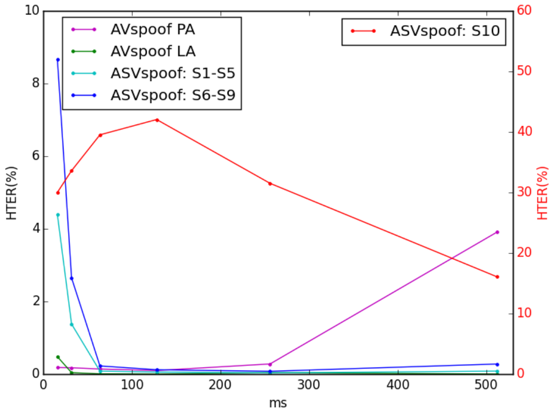


Figure 5: *Impact of frames lengths on the performance of the proposed LDA-based approach, evaluated on the three datasets: ASVspoof, AVspoof-LA and AVspoof-PA.*

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 [42], which also provides high frequency resolution in the low frequency regions and leads to large gains on S10 attack condition.

Finally, the analysis also clearly shows 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 spoofing attacks.

7 conclusions

In one of the first efforts, this paper investigated detection of both physical access attacks and logical access attacks. In this context, we proposed a novel approach that detects presentation attacks based on 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, Interspeech 2015 ASVspoof challenge and AVspoof lead to the following observations:

1. The proposed approach which does not make any speech signal 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 approach 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 show that none of the approaches truly generalize across databases. Such a claim arises despite observing that LFCC based system trained on ASVspoof leads to a low HTER on AVspoof-LA test set because a small modification i.e. by just replacing the triangular shaped filters by rectangular shaped 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 provide benefits such as simple feature extraction with no speech signal related assumption, 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 approaches to learn relevant information or features and the classifier jointly from the raw speech signal with little or no prior knowledge. Such methods of discovering features could lead to better understanding of the problem.

2. The cross-domain studies show that there is need for more resources and further research on how to make the counter-measure systems robust or domain invariant. On the latter aspect, we aim to explore multiple classifier fusion techniques, 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[7] for data collection and development of counter-measures.

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*, 40(3):614–634, March 2001.

[2] Serife Kucur Ergunay, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel. On the vulnerability of speaker verification to realistic voice spoofing. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, September 2015.

[3] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, 66:130–153, 2015.

[4] ISO/IEC JTC 1/SC 37 Biometrics. DIS 30107-1, information technology —– biometrics presentation attack detection. American National Standards Institute, January 2016.

[5] Johnny Mariéthoz and Samy Bengio. Can a professional imitator fool a GMM-based speaker verification system? Technical Report Idiap-RR-61-2005, IDIAP, 2005.

[6] Zhizheng Wu, Tomi Kinnunen, Nicholas W. D. Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proc. of Interspeech*, pages 2037–2041, 2015.

[7] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. In *Proc. of Interspeech*, 2015.

[7] https://www.ntnu.edu/aimt/swan

[8] André Anjos, Laurent El-Shafey, Roy Wallace, Manuel Günther, Christopher McCool, and Sébastien Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *Proc. of the ACM International Conference on Multimedia*, pages 1449–1452. ACM, 2012.

[9] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien 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)*, September 2016.

[10] Pavel Korshunov and Sébastien Marcel. Cross-database evaluation of audio-based spoofing detection systems. In *Proc. of Interspeech*, 2016.

[11] P.L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(8):2280–2290, Oct 2012.

[12] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Synthetic speech detection using temporal modulation feature. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7234–7238, May 2013.

[13] 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*, pages 1–8, Sept 2013.

[14] Federico Alegre, Ravichander Vipperla, Asmaa Amehraye, and Nicholas Evans. A new speaker verification spoofing countermeasure based on local binary patterns. In *Proc. of Interspeech*, page 5p, 2013.

[15] Jakub Gałka, Marcin Grzywacz, and Rafał Samborski. Playback attack detection for text-dependent speaker verification over telephone channels. *Speech Communication*, 67:143 – 153, 2015.

[16] Tanvina B Patel and Hemant 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.

[17] Md Jahangir Alam, Patrick Kenny, Gautam Bhattacharya, and Themos Stafylakis. Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015. In *Proc. of Interspeech*, 2015.

[18] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa. Relative phase information for detecting human speech and spoofed speech. In *Proc. of Interspeech 2015*, 2015.

[19] Yi Liu, Yao Tian, Liang He, Jia Liu, and Michael T Johnson. Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing. *Proc. of Interspeech 2015*, 2:1, 2015.

[20] Phillip L De Leon, Bryan Stewart, and Junichi Yamagishi. Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In *Proc. of Interspeech*, pages 370–373, 2012.

[21] Akio Ogihara, UNNO Hitoshi, and Akira Shiozaki. Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 88(1):280–286, 2005.

[22] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Proc. of Interspeech*, 2015.

[23] Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, and Kai Yu. Robust deep feature for spoofing detection-the SJTU system for ASVspoof 2015 challenge. In *Proc. of Interspeech*, 2015.

[24] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, and Haizhou Li. Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for asvspoof 2015 challenge. In *Proc. of Interspeech*, 2015.

[25] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Spoofing detection from a feature representation perspective. In *Proc. of ICASSP*, pages 2119–2123. IEEE, 2016.

[26] Kristine Tanner, Nelson Roy, Andrea Ash, and Eugene H Buder. Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy? *Journal of Voice*, 19(2):211–222, 2005.

[27] Lindsey K Smith and Alexander M Goberman. Long-time average spectrum in individuals with parkinson disease. *NeuroRehabilitation*, 35(1):77–88, 2014.

[28] Suely Master, Noemi de Biase, Vanessa Pedrosa, and Brasília Maria Chiari. The long-term average spectrum in research and in the clinical practice of speech therapists. *Pró-Fono Revista de Atualização Científica*, 18(1):111–120, 2006.

[29] Elvira Mendoza, Nieves Valencia, Juana Muñoz, and Humberto Trujillo. Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of Voice*, 10(1):59–66, 1997.

[30] Sue Ellen Linville and Jennifer Rens. Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice*, 15(3):323–330, 2001.

[31] T Leino. Long-term average spectrum study on speaking voice quality in male actors. In *Proc. of the Stockholm Music Acoustics Conference*, volume 93, pages 206–210, 1993.

[32] Johan Sundberg. Perception of singing. *The psychology of music*, 1999:171–214, 1999.

[33] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981.

[34] Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147, 1998.

[35] 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*, pages 209–243, 1963.

[36] A. V. Oppenheim and R.W. Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5):95–106, 2004.

[37] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. of ICASSP*, volume 2, pages 1331–1334 vol.2, Apr 1997.

[38] 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*, 28(4):357–366, Aug 1980.

[39] Phu Ngoc Le, Eliathamby Ambikairajah, Julien Epps, Vidhyasaharan Sethu, and Eric H. C. Choi. Investigation of spectral centroid features for cognitive load classification. *Speech Communication*, 53(4):540–551, April 2011.

[40] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, 1986.

[41] Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik, and Vadim Shchemelinin. STC anti-spoofing systems for the ASVspoof 2015 challenge. *arXiv preprint arXiv:1507.08074*, 2015.

[42] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Proc. of Odyssey*, pages 283–290, 2016.

[43] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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*, 27(3):187–207, 1999.

[44] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. of ICASSP*, volume 1, pages 137–140. IEEE, 1992.

[45] John J. Ohala. Respiratory activity in speech. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modeling*. Kluwer Academic Publishers, 1990.