Presentation Attack Detection Using Long-Term Spectral Statistics for Trustworthy Speaker Verification

Hannah Muckenhirn
Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Email: hannah.muckenhirn@idiap.ch

Mathew Magimai-Doss and Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland
Email: {mathew,sebastien.marcel}@idiap.ch

***Abstract*—In recent years, there has been a growing interest in developing countermeasures against non zero-effort attacks for speaker verification systems. Until now, the focus has been on logical access attacks, where the spoofed samples are injected into the system through a software-based process. This paper investigates a more realistic type of attack, referred to as physical access or presentation attacks, where the spoofed samples are presented as input to the microphone. To detect such attacks, we propose a binary classifier based approach that uses long-term spectral statistics as feature input. Experimental studies on the AVspoof database, which contains presentation attacks based on replay, speech synthesis and voice conversion, shows that the proposed approach can yield significantly low detection error rate with a linear classifier (half total error rate of 0.038%). Furthermore, an investigation on Interspeech 2015 ASVspoof challenge dataset shows that it is equally capable of detecting logical access attacks.**

I. Introduction

Automatic Speaker Verification (ASV) systems can achieve a high accuracy in the presence of zero-effort impostors, i.e., speakers that simply attempt to be accepted by the system as another person while using their own voice. However, these systems have been shown to be vulnerable to more elaborated attacks if no countermeasures are implemented [1]. Presentation attacks, also called spoofing attacks, refer to the presentation of falsified or altered samples to a biometric sensor to induce illegitimate acceptance. In this paper, we investigate how to differentiate genuine accesses from the three types of presentation attacks that represent a real threat to ASV systems: replay, voice conversion and speech synthesis.

There has recently been an increasing amount of research on developing countermeasures against attacks to ASV systems, a review of some of them can be found in [2]. A display of this trend is the high participation to the "Automatic Speaker Verification Spoofing and Countermeasures Challenge" [3] during the 2015 edition of Interspeech. The ASVspoof database, used for this challenge, is one of the largest database in attack detection and contains voice conversion and speech synthesis attacks executed via logical access (point 2 in Figure 1). For such attacks, it is assumed that the spoofed samples can directly be injected into the system through a software-based process.

However, according to the ISO standard 30107-1 [4], a presentation attack is performed via physical access, i.e., at the sensor level (point 1 in Figure 1). Indeed, an attacker is unlikely to have access to the system's software. In a more realistic setup, the attacker plays back a recorded utterance to the system. This utterance can either be directly obtained from the real speaker or can be forged with voice conversion or speech synthesis algorithms.

The contributions of this paper are twofold. First, we develop, to the best of our knowledge, the first countermeasures against attacks performed via physical access. To do so, we use the AVspoof database [1]. Secondly, we employ a novel feature representation, which is based on statistics of the log magnitude spectrum and can be easily classified with a linear classifier to detect both physical and logical access attacks.

The remainder of the paper is organized as follows. Section II gives a brief background on Presentation Attack Detection (PAD). Section III then motivates and present the proposed approach of using long-term spectral statistics for PAD. Section IV describes the database and the experimental setup. Section V presents the results and finally Section VI concludes.

II. Background

There are four ways of attacking an ASV system: impersonation, replay, speech synthesis and voice conversion. Impersonation, which refers to the human mimicking of another voice, has been shown to not pose a real threat [5]. Thus, this paper focuses on the three other types of attacks.

Replay is the most feasible attack as the attacker only needs a record and play device. In the literature, research on detection of replay attacks has been investigated and mainly focuses on characteristics related to channel noise and reverberation [6], [7].

For speech synthesis and voice conversion, as mentioned in the introduction, we need to differentiate between attacks performed via logical and physical access, also referred to as indirect and direct attacks, respectively. Logical access attacks correspond to the point 2 in Figure 1. It is assumed that the attackers have hacked into the system to directly inject the spoofed samples. On the other hand, physical access attacks,

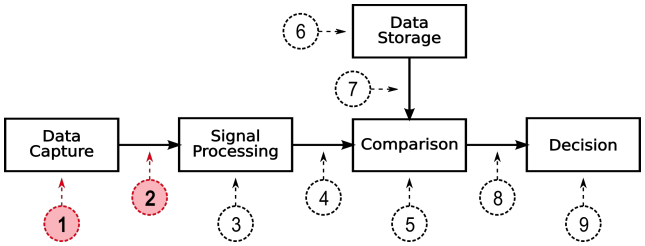

Fig. 1. Potential points of attack in a biometric system, as defined in the ISO-standard 30107-1. Points 1 and 2 correspond respectively to attacks performed via physical access and via logical access.

which correspond to point 1 in Figure 1, are performed at the sensor level. The attackers play the samples directly to the system's microphone. Attacks via physical access are more likely to happen as they require less technical expertise.

There is a fair amount of literature on the detection of logical access attacks with synthetic speech [2]. The methods to detect such attacks, whether generated by voice conversion or speech synthesis algorithms, have mainly focused on the use of features such as the signal phase [8], [9], cepstral coefficients [10]–[12], pitch patterns [13], [14] or the long-term modulation spectrum [15]. There are also approaches that are based on the detection of "pop noise" [16]. However, the lack of databases and standard protocols renders the comparison between these systems difficult [17]. The ASVspoof 2015 Challenge was designed to palliate to this issue by proposing such a common framework. Thus, this challenge is the best source to compare performance for the detection of attacks performed via logical access. Among the 16 teams that participated to this challenge, the one that achieved the best performance used features related to cochlear filter cepstral coefficients, instantaneous frequency and Mel-frequency cepstral coefficients classified with a Gaussian Mixture Model (GMM) [18]. In a more recent work [19], these features were augmented with source-related features such as the fundamental frequency and the strength of excitation, and were found to be beneficial. A performance comparison of 19 features on the ASVspoof database can be found in [20].

However, to the best of our knowledge, there is no research yet on speech synthesis and voice conversion attacks performed via physical access besides a performance study evaluating state-of-the-art PAD systems, which were originally developed for the detection of logical access attacks, on physical access attacks [21].

III. Proposed approach: long-term spectral statistics based PAD

In this section, we first motivate the use of long-term spectral statistics based features for PAD, and then present the proposed approach.

A. Motivation

In automatic speech processing, spectral statistics are employed for various purposes. The Long Term Average Spectrum (LTAS) is used in the clinical domain as a voice quality measurement. It is employed for example for the early detection of voice pathology [22] or Parkinson disease [23] or for evaluating the effect of speech therapy or surgery on the voice quality [24]. In addition to assessing voice quality, LTAS has also been used to investigate voice characteristics. For example, to differentiate between speakers gender [25] and speakers age [26], and also to study singers and actors voices [27], [28].

Voice quality is an informative measure for PAD. The first information it can bring is the channel degradation. For any presentation attack, playing the spoofed sample through loudspeakers could affect the signal quality. In replay attacks, the noise introduced by the microphone during the recording of the original sample could further affect the quality. The second type of information a voice quality measure can bring is the naturalness of the speech. When listening to synthetic speech, one can observe that the speech though intelligible is still far from sounding natural. Indeed, artificially-generated speech introduces some artefacts into the signal. From these two elements, namely, channel degradation and naturalness, we can expect that voice quality related features could yield better discrimination between genuine accesses and presentation attacks, whether it is replay, speech synthesis or voice conversion attacks.

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) [29] and Cepstral Variance Normalization (CVN) [30] to handle channel variability. Formally, the cepstrum is the Fourier transform of the log magnitude spectrum [31], [32]. Thus, the mean and variance of the log magnitude spectrum is indicative of channel variability, which is a desirable feature for PAD.

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. In the following section, we propose an approach along that line.

B. Approach

The proposed approach consists of extracting long-term spectral statistics and using them as feature input to a classifier to detect presentation attacks.

In order to extract long-term spectral statistics, we split the input utterance or speech signal x 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 frequencies, 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. For each frame m, this process yields a vector of DFT coefficients $\mathbf{X}_m = [X_m[0] \cdots X_m[k] \cdots X_m[\frac{N}{2}-1]]^{\mathrm{T}}$.

Next, 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:

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

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

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

This procedure yields one *single* vector representation per utterance, consisting of the mean and standard deviation. The single vector is subsequently fed into a binary classifier to decide if the utterance is a genuine input or an attack. In the present work, we investigate two classifiers: a linear classifier based on Linear Discriminant Analysis (LDA) and a Multi-Layer Perceptron (MLP) with one hidden layer.

IV. Experimental setup

The development of this system is based on the open-source toolbox Bob [33] and QuickNet[1] and all the experiments are reproducible.[2]

A. Databases and protocol

Even though our interest lies in physical access attacks rather than logical access attacks, for the sake of completeness we also present studies on logical access attacks. We use the Audio-Visual Spoofing (AVspoof) database for the physical access attacks study and the Automatic Speaker Verification Spoofing (ASVspoof) database for the logical access attacks study. In the remainder of this section, we describe the two databases along with their specific protocols.

1) AVspoof: The AVspoof database[3] [1] contains replay attacks, as well as speech synthesis and voice conversion attacks both produced via logical and physical access. As explained in the introduction, in this paper we only consider the physical accesses, a.k.a. presentation attacks. 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) and "free speech" (the participants talk freely for 3 to 10 minutes).

Free speech is only used in the training and development phases. However, it is not used when testing the system, as it is not realistic to have someone speaking for 3 to 10 minutes to login into the system.

The attacks are played with four different loudspeakers: the loudspeakers of 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. This enables the database to be more general as different devices do not affect the signal in the same manner.

TABLE I
Number of speakers and utterances for each set of the AVspoof database: train, development and evaluation.

data set	speakers		utterances			
	male	female	genuine	replay	synthesis	conversion
train	10	4	4973	2800	980	34800
development	10	4	4995	2800	980	34800
evaluation	11	5	4376	3200	1120	39000

The data is divided into three subsets, each containing a set of non-overlapping speakers: the training set, the development set and the evaluation set, presented in Table I. The training set is used to optimize the parameters of the classifier. The development set is used to choose the threshold as to obtain an Equal Error Rate (EER), i.e., the false acceptance rate and the false rejection rate are equal. Finally, the evaluation set is used to assess the performance of the system once all the parameters and hyper-parameters values are fixed. We evaluate the performance of our system with the Half Total Error Rate (HTER) computed on the evaluation set.

2) 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. Out of which only 5 types of attacks (S1 to S5) are in the training and development set, while all the 10 types of attacks (S1 to S10) are in the evaluation set. This allows to evaluate the systems on known (S1 to S5) and unknown attacks (S6 to S10). The evaluation protocol involves the estimation of EER independently for each type of attack. 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. The full description of the database and the evaluation protocol is given in [3].

B. System

In this section, we present the details of the PAD system based on the proposed approach.

1) Preprocessing: In the case of presentation attacks (AVspoof dataset), there is an indicative noise at the beginning and end of the utterance. This noise corresponds to the laps of time during which the button to play and stop the sequence is pressed. Even though presence of such noise could be indicative of presentation attacks, we did not want our system to be biased and to rely on these portions to differentiate between real accesses and attacks. To remove these segments at the beginning and the end of the utterance, we used an energy-based Voice Activity Detection (VAD) algorithm. The

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

[2] source code: https://pypi.python.org/pypi/bob.paper.biosig2016

[3] publicly available at https://www.idiap.ch/dataset/avspoof

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

energy values are computed over frames of 20ms with an overlap of 10ms, normalized and then classified into two classes: speech and silence. In the case of logical access attacks (ASVspoof dataset), there is no such needs. So, we did not perform any preprocessing.

2) Feature extraction: The only hyper-parameters in our feature extraction scheme, as defined in Eqn. (2) and Eqn. (3), are: the frame size, the frame shift and the pre-emphasis coefficient applied to each frame to enhance the high frequencies.

The underlying idea of the proposed approach is that the attacks could be detected based on long-term 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 spectral statistics.

In the case of detecting presentation attacks, our interest primarily lies in exploiting the effect of channel on the speech signal. The channel information can be presumed to be spread across different frequencies. Thus, frequency resolution may not be crucial. In other words, conventional short-term speech processing can be sufficient. So, we use parameters values that are very common in speech processing: frame size of 32ms and frame shift of 10ms.

In the case of logical attacks, however, the spoofed speech signal is directly injected into the system and there is no channel effect like for the presentation attacks. So, frequency resolution could be important in this case. We determined the frame size based on cross validation, while keeping the frame shift and pre-emphasis coefficient the same as in the case of presentation attacks. More precisely, we varied the frame size from 32 ms to 512 ms and chose the frame size that yielded the lowest EER on the ASVspoof development data. It was found that frame size of 256 ms yields 0% of EER.

3) Classifier: We investigated two classifiers, namely, a linear classifier based on LDA and a non-linear classifier based on MLP. The input to the classifiers are the spectral statistics estimated at 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: For AVspoof, we use an MLP with one hidden layer composed of 200 units and for ASVspoof with one hidden layer composed of 1000 units. The difference in the number of hidden units is primarily due to the fact that the input feature dimension in the case of ASVspoof (based on DFT of 256 ms) is larger than AVspoof (based on DFT of 32 ms). The MLP classifier was trained using the back propagation algorithm with early stopping criteria.

V. Results

A. Physical access attacks

Table II presents the results for the detection of presentation attacks on the evaluation set of the AVspoof database in terms of HTER as well as the number of misclassified attacks and genuine accesses to give further insights into the system performance.

TABLE II
Performance on the evaluation set of the AVspoof database.

feature	algorithm	HTER	misclassified attacks / genuine
μ	LDA	0.263%	10 / 22
	MLP	0.057%	20 / 3
σ	LDA	2.753%	831 / 157
	MLP	1.927%	383 / 130
$[\mu, \sigma]$	LDA	0.038%	13 / 2
	MLP	0.049%	23 / 2

We observe that, whether the classification is done with a MLP or a LDA classifier, the long term spectral average clearly outperforms the standard deviation. However, the fusion of the two features lowers the error rate. When we compare the performance of using the feature-level fusion of the mean and standard deviation with a MLP or with a LDA, we see that the LDA classifier slightly outperforms the MLP. This shows that the features based on long term spectral statistics are highly discriminative and can be simply classified using a linear classifier to detect presentation attacks.

In [21], the authors have benchmarked state-of-the-art systems, which were originally proposed for the detection of logical access attacks, on the AVspoof database. The lowest HTER obtained on the evaluation set was 2.70%. This was achieved by extracting 20 dimensional Rectangular Filter Cepstral Coefficients with their first and second derivatives from short-term speech signal and using a GMM of 512 components to classify these features. It can be observed that all the proposed systems, i.e., μ only, σ only and $[\mu, \sigma]$, yield better performance than that.

B. Comparative study on logical access attacks

Table III compares the systems based on our approach against the five best systems (denoted as System ID A-E) proposed in the ASVspoof 2015 challenge [3]. The ASVspoof 2015 challenge systems typically employed multiple features and fusion techniques. For example, the team that achieved the best performance [18] used a fusion of cochlear filter cepstral coefficients, instantaneous frequency and Mel-frequency cepstral coefficients, classified with a GMM. Similarly, the second best system [34] employed fusion of multiple features based on Mel-frequency cepstrum and phase spectrum; transforming them into i-vectors; and finally classifying the i-vectors with a support vector machine. More information can be found in the respective citations provided in the table.

We built two systems using $[\mu,\ \sigma]$ as the feature input, namely, LDA-based and MLP-based. On average, both systems perform better than the ones proposed in the ASVspoof 2015 challenge. It can be observed that the LDA-based approach yields one of the lowest error rates for both Known and Unknown attacks scenario. However, the MLP-based approach yields a higher error rate on the Known attacks scenario and a lower one on the Unknown attacks scenario.

TABLE III
PERFORMANCE ON THE ASVSPOOF 2015 CHALLENGE DATA SET IN TERMS OF EER. FOR COMPARISON PURPOSES, FIVE BEST PERFORMANCES INDICATED AS SYSTEMS A-E WERE TAKEN FROM [3].

System ID	Equal Error Rates (EERs)		
	Known attacks	Unknown attacks	Average
A [18]	0.408	2.013	1.211
B [34]	0.008	3.922	1.965
C [35]	0.058	4.998	2.528
D [36]	**0.003**	5.231	2.617
E [37]	0.041	5.347	2.694
Proposed approach: LDA	0.026	2.086	1.056
Proposed approach: MLP	0.270	**0.781**	**0.525**

TABLE IV
PERFORMANCE OF THE PROPOSED APPROACH ON EACH ATTACK IN TERMS OF EER.

Known attacks	S1	S2	S3	S4	S5
LDA	0.000	0.043	0.000	0.000	0.086
MLP	0.043	0.234	0.022	0.032	1.019
Unknown attacks	**S6**	**S7**	**S8**	**S9**	**S10**
LDA	0.086	0.030	0.083	0.032	10.197
MLP	0.744	0.118	0.054	0.202	2.787

Table IV presents the performance of our approach for each type of attack. We can observe a trend similar to AVspoof, i.e., the LDA classifier based system outperforms MLP classifier based system for all attacks except for S10, which is still an issue for most of the approaches proposed in the literature.

VI. DISCUSSION AND CONCLUSION

This paper proposed an approach to detect non zero-effort attacks on speaker verification systems based on long-term spectral statistics. Even though the main focus of this paper was on physical access attacks, we also studied logical access attack detection in order to relate to existing works. Our investigations showed that the proposed approach yields a very low error rate for both types of attacks and outperforms existing systems. Specifically, for physical access attacks, the approach achieves a HTER of 0.038% using a LDA and 0.049% using a MLP on the purpose-built AVspoof database. For the detection of logical access attacks, we used the ASVspoof 2015 Challenge data set and obtained an average EER of 1.056% with LDA classifier based system and 0.525% with MLP classifier based system.

Although the proposed approach scales well to both types of attacks, the frame size used to compute the spectral statistics is not the same: 32ms for physical attacks, 256ms for logical attacks. Our original reasoning was that frequency resolution may not matter for presentation attacks but may matter for logical access attacks. In order to ascertain that, we analyzed the LDA coefficients estimated for the different frequency bins. We found that in the case of physical attacks, importance is given to all frequency bins. This supports our assumption that in presentation attacks, the channel effect will impact the whole bandwidth, thus the relevant information is spread across different frequency bins. However, in the case of logical access attacks, importance is given to a few frequency bins that are well below 50 Hz, i.e., discriminative information in the frequency domain is highly localized. Thus, we consequently need a higher frequency resolution.

To conclude, our study showed that physical and logical attacks can be effectively detected using simple utterance level long-term spectral statistics combined with a linear classifier. This is particularly promising when considering the existing approaches, which rely on complex feature extraction and classifiers.

Our future work will focus on further investigating the generalization capabilities of the proposed approach under different mismatched conditions. For instance, training either on presentation attacks or logical access attacks and evaluating on the other.

ACKNOWLEDGMENT

This work was funded by the SNSF project "Unified Speech Processing Framework for Trustworthy Speaker Recognition (UniTS)" and by the Norwegian SWAN project.

REFERENCES

[1] S. K. 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*, Sep. 2015.

[2] 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.

[3] Z. Wu, T. Kinnunen, N. 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.

[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, Idiap-RR Idiap-RR-61-2005, 2005.

[6] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*. Springer, 2011, pp. 274–285.

[7] Z. Wang, G. Wei, and Q. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 4. IEEE, 2011, pp. 1708–1713.

[8] P. D. 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, 2012.

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

[10] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[11] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[12] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 283–290.

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

[14] 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.

[15] Z. Wu, X. Xiao, E. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7234–7238.

[16] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 259–263.

[17] N. Evans, J. Yamagishi, and T. Kinnunen, "Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics," *IEEE Speech and Language Technical Committee Newsletter*, 2013.

[18] T. Patel and H. 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.

[19] ——, "Effectiveness of fundamental frequency (F0) and strength of excitation (SOE) for spoofed speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5105–5109.

[20] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. of Interspeech*, 2015.

[21] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Proc. of Interspeech (to appear)*, 2016.

[22] K. Tanner, N. Roy, A. Ash, and E. 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.

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

[24] S. Master, N. D. Biase, V. Pedrosa, and B. 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.

[25] 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.

[26] S. 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.

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

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

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

[30] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.

[31] 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, pp. 209–243.

[32] 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.

[33] 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 20th ACM international conference on Multimedia*, 2012, pp. 1449–1452.

[34] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the asvspoof 2015 challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5475–5479.

[35] 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.

[36] X. Xiao, X. Tian, S. Du, H. Xu, E. 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.

[37] M. 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.