Abstract—Biometric systems are nowadays employed across a broad range of applications. They provide high security and efficiency and, in many cases, are user friendly. Despite these and other advantages, biometric systems in general and Automatic Speaker Verification (ASV) systems in particular can be vulnerable to attack presentations. The most recent ASVSpoof 2019 competition showed that most forms of attacks can be detected reliably with ensemble classifier-based presentation attack detection (PAD) approaches. These, though, depend fundamentally upon the complementarity of systems in the ensemble. With the motivation to increase the generalisability of PAD solutions, this paper reports our exploration of texture descriptors applied to the analysis of speech spectrogram images. In particular, we propose a common fisher vector feature space based on a generative model. Experimental results show the soundness of our approach: at most, 16 in 100 bona fide presentations are rejected whereas only one in 100 attack presentations are accepted.

I. INTRODUCTION

In the last decades, biometric systems have experienced a broad development since they provide high security, efficiency, and in many cases they are more user friendly than the traditional credential-based access control systems. In spite of those advantages, they are vulnerable to attack presentations, which can be easily carried out by a non-authorised individual without having enough computational knowledge [1]. Those, in which biometric systems, are commonly deployed are in turn employed to gain access to several applications such as bank accounts, to unlock smartphones, or to circumvent a border control. Specifically, Automatic Speaker Verification (ASV) systems have shown a verification performance deterioration when different Presentation Attack Instruments (PAIs) such as replay, voice conversion, speech synthesis and impersonation are launched [2].

In order to deal with those security threats, several Presentation Attack Detection (PAD) approaches have been proposed. These methods try to determine whether a given sample stems from a live subject (i.e., it is a bona fide presentation, BP) or from a synthetic replica (i.e., it is an attack presentation, AP). According to the last ASV Spoof 2019 competition [3], the top Deep Neural Networks (DNN)-based PAD techniques have achieved a remarkable detection performance to spot PAI species which can be either known or unknown a priori in training time. Despite the promising results reported, those methods are time-consuming, hence being unsuitable for lightweight security applications, e.g., for smartphones.

To overcome the aforementioned drawbacks, very few works have explored texture-based representation for audio PAD. Alegre et al. [2] proposed a PAD method based on the combination of Local Binary Patterns (LBP) and one-class classifiers. Even if the proposed algorithm reported a poor detection performance for some unknown attacks such as voice conversion, this work showed the generalisation capability of the proposed texture-based representation for audio PAD.

Motivated by that fact, we explore in this work several image processing texture descriptors in combination with a Support Vector Machine (SVM), which have been successfully employed for fingerprint [4] and face [5] PAD. SVMs are popular as classifiers since they perform well in high-dimensional spaces and avoid over-fitting. In order to improve the generalisation capability of the analysed texture descriptors, we also utilise the Fisher Vector (FV) representation, which has shown remarkable results for unknown fingerprint [6] and face [7] PAD. By assuming that the unknown attacks share more texture, shape, and appearance features with known PAI species than with BP samples, the FV representation defines a common feature space from the parameters learned by an unsupervised Gaussian Mixture Model (GMM) in order to deal with the generalisation to unknown attacks.

The remainder of this paper is organised as follows. The speech-to-image domain transformation techniques are presented in Sect. [II]. Sect. [III] briefly describes the texture descriptors analysed in this work. Sect. [IV] describes the experimental protocol and Sect. [V] presents the results. Finally, conclusions and future work directions are presented in Sect. [VI].

II. FROM SPEECH TO IMAGES DOMAIN

The visualisation of audio/speech signals is key to many audio analysis tasks, usually involving: (i) time-domain,
shown to be beneficial for ASV and utterance verification [13]. Features were designed specifically for PAD but have also

tion more closely related to that of human perception. These
cQTC representations. CQCCs offer a time-frequency resolu-
tion for lower frequencies and a greater temporal
resolution for higher frequencies.
The CQT is a perceptually motivated approach to time-
frequency analysis. In contrast to Fourier-based approach es,
さらには時系列でのこれらの実験が行われた。この研究は、特に時系列データの解析に適しており、時系列データをより適切に解析するための新しいアプローチを提供している。

(i) frequency-domain, or (ii) time-frequency-domain represen-
tations known as spectrograms, which show the signal
amplitude over time at a set of discrete frequencies. Many
time-frequency representations have been proposed, each with
different characteristics. Keeping in mind that an audio signal
can be represented as an image, as shown in Fig. 1 in this
work we focus on the following four time-frequency represen-
tations: (i) short-time Fourier transform (STFT) [8], (ii) linear
frequency cepstral coefficients (LFCC) [9], (iii) constant Q
transform (CQT) [10], and (iv) constant Q cepstral coefficients
(CQCC) [11], [12].

The STFT is a time-frequency decomposition based upon
the application of Fourier analysis to short segments or win-
dows of the audio signal. As such, it is effectively a filter bank
where the bandwidth of each filter is constant and is related
to the window function.

LFCC coefficients are computed from the STFT by applying
the discrete cosine transform (DCT) [8]. Generally, only lower-
order coefficients are retained since they represent the vocal
tract configuration.

The CQT is a perceptually motivated approach to time-
frequency analysis. In contrast to Fourier-based approaches,
the bin frequencies of the filterbank are geometrically dis-
tributed. Compared to the STFT, the CQT has a greater fre-
quency resolution for lower frequencies and a greater temporal
resolution for higher frequencies.

CQCCs stem from the application of cepstral processing to
CQT representations. CQCCs offer a time-frequency resolu-
tion more closely related to that of human perception. These
features were designed specifically for PAD but have also
shown to be beneficial for ASV and utterance verification [13].

III. TEXTURE ANALYSIS BASED APPROACHES

In order to analyse the texture features extracted from the
time-frequency image representations described in Sect. II
we selected several well-known texture descriptors, which
have been successfully employed for the PAD task across
different modalities such as fingerprint [4] and face [5]: Local
Binary Patterns (LBP) [14], Multi-block LBP (MB-LBP) [15],
Local Phase Quantisation (LPQ) [16], and Binarized Statistical
Image Features (BSIF) [17].

A. Local Binary Patterns

LBP [14] represents an image as a histogram of uniform
patterns which allows capturing shape and texture features
from the image. Specifically, given a circular image patch X
with radii σ and N pixels around the center, the LBP descriptor
is defined over X as:

\[ \text{LBP}_{N,\sigma} = \sum_{i=0}^{N-1} f(g_i - g_c)2^i, \]  

where \( g_i \) with \( i = 0 \ldots N - 1 \) are gray intensity values around
the center \( g_c \). In addition, \( f(g_i - g_c) \) is computed as:

\[ f(g_i - g_c) = \begin{cases} 
1, & g_i - g_c \geq 0 \\
0, & g_i - g_c < 0 \end{cases} \]  

In order to capture more information and thereby increase
the descriptor distinctiveness, we compute the LBP patterns
for various radii \( \sigma \) and number of neighbors \( N \) (i.e., \( \sigma = \{1, 2, 3\} \) and \( N = \{8, 16, 24\} \)), as in [6].

B. Multi-block Local Binary Patterns

The MB-LBP [15] encodes the intensities of rectangular
regions with the LBP operator, which allows describing several
local structures of an image. Whereas the LBP descriptor
in [14] is defined for each pixel by thresholding the 3 × 3
neighbourhood pixel values with the center pixel value,
the MB-LBP operator represents each pixel \( x \) by comparing
the central rectangle average intensity \( g_x \) with those of its
neighbourhood rectangles. Therefore, it can detect numerous
image structures such as lines, edges, spots, flat areas, and
corners [15], at different scales and locations. Unlike LBP,
the MB-LBP descriptor can thus capture large scale structures
that may be the dominant features of images, with 256 binary
patterns. In our work, we compute the MB-LBP descriptor for
several rectangle sizes \( R_x = \{3, 5, 7, 9\} \).

C. Local Phase Quantization

LPQ [16] is a texture descriptor designed to deal with
blurred images. It represents an image patch of size \( l \times l \)
centred on a pixel \( x \) as a 256-histogram by using the local
phase information, extracted by a STFT. Let \( F_{u_{i=1..4}} \) be the
outputs of the STFT for the pixel \( x \) using four bi-dimensional
spatial frequency \( u_0, u_1, u_2 \) and \( u_3 \), the LPQ features for
\( x \) are defined as a vector whose components are formed by
stacking the real and imaginary part of \( F_{u_{i=1..4}} \). Subsequently,
the vector elements are quantized using a previously defined
function and then represented as a integer value in the range
[0 . . . 255]. In order to make the LPQ coefficients statistically
independent, a decorrelation step based on whitening trans-
form was performed.

Fig. 1. A speech sample together with its texture image representation.
D. Binarized Statistical Image Features

BSIF \cite{17} is a local descriptor based on Independent Component Analysis (ICA), which uses pre-learned filters to obtain statistically meaningful representation of the data. Let $X$ be an image patch of size $l \times l$ and $W = \{W_1, \ldots, W_N\}$ a set of linear filters of the same size as $X$. Then, we can compute the binarised responses $b_i$, with $i = 1, \ldots, N$, as:

$$b_i = \begin{cases} 
1 & \sum_{u,v} W_i(u,v)X(u,v) > 0 \\
0 & \text{otherwise}
\end{cases} \quad (3)$$

All the filter responses $b_i$ are subsequently stacked to form a single bit string $b$ of size $N$ per image pixel. Consequently, that bit string is transformed in a decimal value, and then a $2^N$ histogram for $X$ is computed. This histogram is finally represented as a 128-component vector \cite{17}. In our work, we adopt 60 filter sets \cite{17} with different sizes $l = \{3, 5, 7, 9, 11, 13, 15, 17\}$ and number of filters $N = \{5, 6, 7, 8, 9, 10, 11, 12\}$ to compute the BSIF descriptors.

E. Fisher Vector encoding

For FV encoding, we build upon the idea proposed by \cite{18, 19}: a Gaussian Mixture Model (GMM) is used to unsupervisedly learn embedding features from a texture descriptor. Generally, a GMM model based on $K$-components, which are represented by their mixture weights ($\pi_k$), Gaussian means ($\mu_k$), and covariance matrix ($\Sigma_k$), with $k = 1, \ldots, K$, allows discovering semantic sub-groups from known PAI and BP samples, which could successfully generalise to and hence enhance the detection of unknown PAI species. Given the parameters learned by a GMM model, the FV computes the average first-order and second-order statistics differences between the texture features and each GMM Gaussian, thereby yielding a $2Kd$ dimensional vector ($d$ is the texture feature dimension). The FV gradient computation between texture features extracted from an given unknown sample and the GMM components represents how its distribution differs from the distribution of the model. Therefore, a better data distribution will reveal more discriminative FV common feature space to detect unknown attacks.

IV. EXPERIMENTAL EVALUATION

The experimental evaluation follows a threefold objective: i) analyse the detection performance of the texture descriptors for known and unknown PAD; ii) establish a benchmark against the speech-based state-of-the-art PAD techniques; and iii) evaluate the computational cost of our proposal.

A. Databases

The experimental evaluation was conducted over the freely available ASVspoof 2019 database \cite{3, 20}, which comprises two assessment scenarios: Logical Access (LAc) and Physical Access (PAc). Both LAc and PAc databases are partitioned in three disjoint datasets: training, development, and evaluation. Whereas the PAIs in the training and development datasets were built with the same algorithms and capture conditions (i.e., it is the known attack scenario), PAIs for the evaluation dataset were generated with different techniques and capture conditions (i.e., it is the unknown attack scenario).

The LAc partition contains PAI samples which were generated using 14 different text-to-speech synthesis (TTS) and voice conversion (VC) technologies: six were designated for known attack assessment (i.e., A01-A06) and 11 for unknown attack (i.e., A07-019 with exception of A16 = A04 and A19 = A06 and hence both two attacks are in the training set). In order to analyse and improve the ASV reliability in different acoustic environments and replay setups, the training and development data for the PAc scenario is created under 27 different acoustic and 9 replay configurations. The replay settings comprise 3 attacker-to-talker (i.e., A, B, C) recording distances and 3 loudspeaker quality (i.e., A, B, C). The evaluation dataset is generated in the same manner as training and development data but with different random acoustic and replay configurations.

B. Baselines and metrics

In order to establish a fair benchmark, we adopt two PAD baseline approaches, which use GMM back-end classifier with either CQCC (B01) or LFCC (B02) features. It should be noted that whereas the baseline approaches employed a bi-cluster GMM model for a BP or AP classification, our analysed FV approach uses it as generative model to fit the BP and AP data distribution. Based on that fact, we evaluated several numbers of Gaussian clusters (i.e., $K = \{64, 128, 256, 512, 1024\}$) for the FV representation.

The experimental evaluation is conducted in compliance with ISO/IEC 30107-Part 3 \cite{11}, analysing two metrics: i) Attack Presentation Classification Error Rate (APCER) which is defined as the proportion of APs wrongly classified as BPs; and ii) Bona Fide Presentation Classification Error Rate (BPCER): which is defined as the proportion of BPs misclassified as APs. In addition, we report the Detection Error Trade-off (DET) curves between both errors, the BPCER values for several security thresholds (i.e., BPCER10, BPCER20, BPCER100), and the D-EER. Finally, to evaluate the performance between the proposed PAD approaches and a ASV system, we report the minimum normalised tandem detection cost function (t-DCF) \cite{21}, which is the primary metric used for the ASVspoof 2019 challenge\cite{7}.

V. RESULTS AND DISCUSSION

A. Known attacks

In order to analyse the detection performance, two sets of experiments were carried out. In the first experiment set, we optimise the detection performance of our texture descriptors in terms of the D-EER for different parameter configurations. Table \cite{11} shows the D-EER for the best parameter setting over the development set in the LAc and PAc scenarios. As it may be observed, among all speech-to-image domain transformations, the CQT reports the best detection performance, thereby resulting a mean D-EER of 6.14% and 5.80% for the LAc

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Method & D-EER (LAc) & D-EER (PAc) \\
\hline
Baseline & \ldots & \ldots \\
CQT & \ldots & \ldots \\
BSIF & \ldots & \ldots \\
FV & \ldots & \ldots \\
\hline
\end{tabular}
\caption{Experimental Results}
\end{table}

\footnote{\url{https://www.asvspoof.org/}}
and PAc scenarios, respectively. In addition, among the texture descriptors, the BSIF unveils the best texture features for the audio PAD task: D-EERs of 0.86% and 4.54% are attained for the LAc and PAc with CQT, respectively, thereby showing its suitability for the audio PAD task.

In a second set of experiments, we evaluate our FV approach for the best texture descriptor (i.e., BSIF) for known attack detection (dev set). Table [I] shows the FV detection performance for the best number of Gaussian clusters per speech-to-image domain transformation. We can first observe that STFT achieves the same detection performance for the LAc scenario for the pooled database (i.e., all PAl species) than the one reported by the single BSIF descriptor: a D-EER of 0.86%, which is approximately three times lower than the one attained by the FV encoding with the images in the CQT domain. Alternatively, a different detection performance is reported for the PAc scenario where a D-EER of 3.68% for the CQT outperforms the one attained by STFT (i.e., 11.05%). Finally, it may be noted that the t-DCF values for both the STFT on LAc and CQT on PAc are respectively below 0.05 and 0.14, hence indicating that the FV provides a high security against PAl to the ASV systems.

B. Unknown attacks

As mentioned in Sect. [IV], one of the goals of this work is to analyse traditional texture descriptors for unknown attack detection. To that end, we select the evaluation dataset and assess the detection performance for the adopted texture descriptors by setting up the same parameters reported for the known attack experiment.

The corresponding results are presented in Tab. [III] and Tab. [IV]. We can note that the BSIF descriptor attains again the best detection performance for most speech-to-image domain transformations: a D-EER of 4.94% for PAc, which is close to the one reported by the known attack scenario (i.e., 3.68%). In addition, the MB-LBP outperforms the remaining descriptors for the STFT-LAc scenario, achieving a D-EER of 10.69%.

Based on this fact, we also evaluated the combination between BSIF and FV for each particular PAl species for the LAc and PAc scenarios in Tab. [IV]. As it can be observed, the CQT achieves the best detection performance for the entire set of LAc-PAl species (i.e., a D-EER of 6.83%). In addition, this outperforms the adopted baselines for the most challenging PAl species for LAc scenario under the ASVSpoof 2019 challenge [4]: D-EERs of 7.91% and 1.94% are respectively achieved for A10 and A13, which are two and five times lower than the ones reported by the baselines. Moreover, their corresponding t-DCF values are better than the ones attained by the baselines.

Consequently, for the Physical Access scenario our CQT-based FV approach attains for the pooled a D-EER of 3.66%, which is three times lower than the one yielded by the baselines (i.e., a D-EER of 11.04% for B01 and a D-EER of 13.54% for B02). Furthermore, we outperform the baselines for most PAl species: a D-EER in the range 11.0-7.64% together with a t-DCF between 0.03-0.20% unveils a reliable and secure generalisation capability for this scenario.

Finally, a tSNE visualisation in Fig. [2] shows that most PAl species share more homogeneous features with each other than with those bona fide presentations. However, the overlap of some PAl species such as A17, A18, AA, BA, and CA with the BP features indicates that the data distribution learned by a GMM using the BSIF features needs to be improved in order to get a better generalisable FV common feature space.
In order to determine whether the FV encoding together with BSIF is suitable to yield a secure ASV system, we evaluated it for several operating points according to the ISO [1]. Fig. 3 shows the DET curves as well as the BPCER values for several security thresholds for the LAc and PAC scenarios over the development and evaluation datasets. As it should be noted, different behaviours are observed for both the known and unknown attack detection. Whereas the FV encoding outperforms only one out of two baselines for unknown attacks, which is marginally worse than the one computed for the known attack detection (i.e., 15.24%). Based on these results, we confirm that the texture characteristics provided by the CQT domain transformation, captured by the BSIF descriptors, and finally encoded by the FV, yields a secure and convenient ASV system.

C. In depth performance analysis

On the other hand, for the unknown attack detection (i.e., eval set), we can see that, the FV encoding is better than both baselines: it improves B01 and B02 by a relative 30.46% for LAc and 81.67% for PAC, thereby confirming the results shown in Tab. IV and hence its soundness to detect this frequent type of threat. Finally, it is worth noting that the combination of BSIF together with the FV representation reports a similar behaviour for known and unknown attacks over the PAC scenario (i.e., gray vs. black on the right DET figure). In particular, for a high security threshold (i.e., an APCER of 1%), the FV attains a BCPER value of 15.56% for unknown attacks, which is marginally worse than the one computed for the known attack detection (i.e., 15.24%). Based on these results, we confirm that the texture characteristics provided by the CQT domain transformation, captured by the BSIF descriptors, and finally encoded by the FV, yields a secure and convenient ASV system.

D. Computational efficiency

In order to evaluate the computational efficiency, we selected the FV encoding for the best parameter configuration (i.e., CQT) and studied its efficiency for several GMM parameter configurations. In the experiments, a PC with an Intel Core i7-8750H processor at 2.2GHz, 16GB RAM was employed. The experimental results report that the FV encoding for $K = 512$ clusters is able to detect an attack presentation attempt in 1.6 sec approximately for CQT samples with a large size...
outperforming the current audio-based baselines. In addition, the FV approach achieved a relative performance improvement of 57.66% for the unknown attack detection, thereby demonstrating that the FV approach is suitable to build a real-time audio PAD system.

VI. CONCLUSION

In this work, we studied the reliability of image texture analysis to develop a secure ASV system. To that end, four traditional texture descriptors which have been widely employed for several PAD tasks were selected. In addition, an application of the well-known FV encoding to audio PAD was proposed. This representation allows the definition of semantic groups given a generative model, such as a GMM.

Experimental results over the ASVSpoof 2019 database confirmed the soundness of the proposal to detect both the known and unknown audio attacks. In particular, the texture characteristics provided by the CQT domain transformation and captured by the BSIF descriptor, reported a D-EER of 16.13% and 5.97% for the LAc and PAc scenarios, respectively, in the presence of unknown attacks. These BSIF descriptors represented in turn that the FV approach is suitable to build a real-time audio PAD system.

Fig. 3. DET curves for the best FV parameter configuration against the baselines for the LAc and PAc scenarios.

As future work research direction, we will investigate a new generative model to tackle the overlapping issues depicted in Fig. 2 and hence improve the generalisation capability of the FV representation. Furthermore, one natural extension of the work would be to look for fusion strategies among methods that are complementary to texture-based PAD.

REFERENCES

[1] ISO/IEC JTC1 SC37 Biometrics, ISO/IEC FDIS 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting, International Organization for Standardization, 2017.

[2] F. Alegre, A. Amehaye, and N. Evans, “A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns,” in Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS), 2013, pp. 1–8.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, “Avspooﬁng 2019: Future horizons in spoofed and fake audio detection,” Proc. Interspeech, 2019.

[4] E. Marasco and A. Ross, “A survey on antispoofing schemes for fingerprint recognition systems,” ACM Computing Surveys (CSUR), vol. 47, no. 2, pp. 1–36, 2014.

[5] J. Galbally, S. Marcel, and J. Fierrez, “Biometric antispoofing methods: A survey in face recognition,” IEEE Access, vol. 2, pp. 1530–1552, 2014.

[6] L. J. González-Soler, M. Gomez-Barrero, L. Chang, A. Pérez-Suárez, and C. Busch, “On the impact of different fabrication materials on ﬁngerprint presentation attack detection,” in Proc. Int. Conf. on Biometrics (ICB), 2019, pp. 1–6.

[7] L. J. González-Soler, M. Gomez-Barrero, and C. Busch, “Fisher vector encoding of Dense-BSIF features for unknown face presentation attack detection,” in Proc. Int. Conf. of the Special Interest Group on Biometrics (Biosig 2020), ser. LINL. GI, September 2020, pp. 1–11.

[8] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-time Signal Processing (2Nd Ed.). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[9] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, “Spoofing attack detection using the non-linear fusion of sub-band classiﬁers,” in Proc. Interspeech, 2020.

[10] J. Brown, “Calculation of a constant Q spectral transform,” Journal of the Acoustical Society of America, vol. 89, no. 1, pp. 425–434, 1991.

[11] M. Todisco, H. Delgado, and N. Evans, “Constant Q cepstral coefﬁcients: A spoofing countermeasure for automatic speaker veriﬁcation,” Computer Speech & Language, vol. 45, pp. 516–535, 2017.

[12] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, “An explainability study of the constant Q cepstral coefﬁcient spoofing countermeasure for automatic speaker veriﬁcation,” in Proc. Odyssey, 2020.

[13] M. Todisco, H. Delgado, and N. Evans, “Articulation rate ﬁltering of CQCC features for automatic speaker veriﬁcation,” in Proc. Interspeech, 2016.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classiﬁcation with local binary patterns,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[15] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, “Face detection based on multi-block lbp representation,” in Proc. Int. Conf. on Biometrics (ICB), 2007, pp. 11–18.

[16] V. Ojansivu and J. Heikilä, “Blur insensitive texture classiﬁcation using local phase quantization,” in Proc. Int. Conf. on Image and Signal Processing (ISP), 2008, pp. 236–243.

[17] J. Kannala and E. Rahtu, “BSIF: Binarized statistical image features,” in Proc. Int. Conf. on Pattern Recognition (ICPR), 2012, pp. 1363–1366.

[18] L. J. González-Soler, M. Gomez-Barrero, L. Chang, A. Pérez-Suárez, and C. Busch, “Fingerprint presentation attack detection based on local features encoding for unknown attacks,” arXiv preprint arXiv:1908.10163, 2019.

[19] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classiﬁcation,” in Proc. European Conf. on Computer Vision (ECCV), 2010, pp. 143–156.

[20] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, and K. Lee, “Avspooﬁng 2019: A large-scale public database of synthesized, converted and replayed speech,” Computer Speech & Language, vol. 64, pp. 101–114, 2020.

[21] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. Reynolds, “Tandem assessment of spoofing countermeasures and automatic speaker veriﬁcation: Fundamentals,” IEEE Trans. on Audio, Speech, and Language Processing, 2020.