proceedings

MDPI

Proceedings

On the Use of Fisher Vector Encoding for Voice Spoofing Detection †

Jahangir Alam

Computer Research Institute of Montreal (CRIM), Montréal, QC H3N 1M3, Canada; Jahangir.Alam@crim.ca

* Correspondence: Jahangir.Alam@crim.ca; Tel.: +1514-840-1235

† Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

Published: 20 November 2019

Abstract: Recently, the vulnerability of automatic speaker recognition systems to spoofing attacks has received significant interest among researchers. A robust speaker recognition system demands not only high recognition accuracy but also robustness to spoofing attacks. Several spoofing and countermeasure challenges have been organized to draw attention to this problem among the speaker recognition communities. Low-level descriptors designed to detect artifacts in spoofed speech are found to be the most effective countermeasures against spoofing attacks. In this work, we used Fisher vector encoding of low-level descriptors extracted from speech signals. The idea behind Fisher vector encoding is to determine the amount of change induced by the descriptors of the signal on a background probability model which is typically a Gaussian mixture model. The Fisher vector encodes the amount of change of the model parameters to optimally fit the new- coming data. For performance evaluation of the proposed approach we carried out spoofing detection experiments on the 2015 edition of automatic speaker verification spoofing and countermeasure challenge (ASVspoof2015) and report results on the evaluation set. As baseline systems, we used the standard Gaussian mixture model and i-vector/PLDA paradigms. For a fair comparison, in all systems, Constant Q cepstral coefficient (CQCC) features were used as low-level descriptors. With the Fisher vector-based approach, we achieved an equal error rate (EER) of 0.1145% on the known attacks, 1.223% on the unknown attacks, and 0.668% on the average. Moreover, with a single decision threshold this approach yielded an EER of 1.05% on the evaluation set.

Keywords: spoofing detection; CQCC; Fisher vector; GMM; i-vector; PLDA

1. Introduction

In a spoofing attack, a person or computer program tries to impersonate the legitimate user of an authentication system. Some examples of voice spoofing attacks are impersonation, replay, speech synthesis, and voice conversion. Replay attacks are also known as presentation or physical access attacks. The remaining three attacks are called logical access attacks [1–5]. With the increasing influence of smart devices in our daily lives, the use of biometric systems is becoming popular in various applications such as phone unlock, access control, smart home assistance, and security. Biometrics, such as voice, face, fingerprint, and iris, are widely utilized for person authentication due to their intrinsic distinctiveness and convenience to use. Voice, as one of the most popular modalities, has received increasing attention in academia and industry in recent years. While speaker recognition systems gain popularity, fraudsters present various voice spoofing attacks to the system and attempt to gain illegitimate access to the authentication system. This problem has attracted the interest of both speech synthesis and speaker recognition researchers. With recent advances in speech synthesis and voice conversion approaches based on deep neural networks (for example, end-to-end direct

waveform modeling, generative adversarial networks (GANS)) [6–9] and combined with the availability of open-source toolkits online, spoofing attacks generated by speech synthesis and voice conversion techniques are potentially more serious [1–4]. A robust speaker recognition system demands not only high recognition accuracy but also robustness to spoofing attacks to provide users with secure and convenient ways to access their personal information. Hence, voice anti-spoofing is crucial to prevent automatic speaker recognition systems from a security breach. Figure 1 shows a voice anti-spoofing system incorporated with an automatic speaker verification (ASV) system. The speech signal from the claimed identity (can be genuine or spoofed) is first passed through an ASV system for verification. If found to be non-target, the claimed identity is rejected. If accepted by the ASV system, then the speech signal is passed through the spoofing detection system to make sure that the claimed speech is genuine. The claimed identity is rejected if it is detected as spoof by the spoofing detection system.

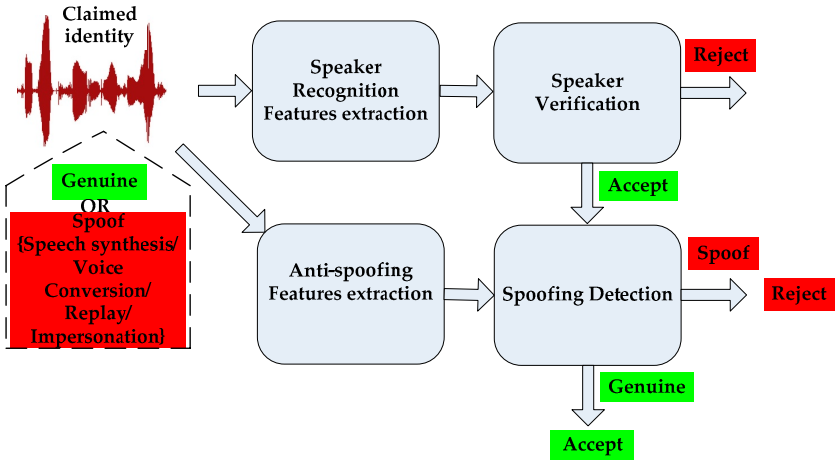

Figure 1. Spoofing detection system incorporated with an automatic speaker verification framework. The speech signal from the claimed identity is passed through the spoofing detection system only when it is accepted by the speaker verification system.

Extraction of anti-spoofing features or countermeasures play a key role in the detection of spoofing attacks. Various low-level descriptors have been developed and investigated as countermeasures since the susceptibility of voice biometrics to spoofing attacks was recognized by the research community. During and after the ASVspoof2015 challenge, many countermeasures based on spectral amplitude, phase [5,10–17], combined amplitude-phase [11–13], and tandem [11,12] have been used for spoofing detection. In ref. [10], the constant Q transform-based cepstral coefficient (CQCC) countermeasure was proposed and evaluated on the first edition of the automatic speaker verification spoofing and countermeasures challenge (ASVspoof2015) corpus. Infinite impulse response-constant Q transform (IIR-CQT)-based cepstral coefficient (ICQC) features were investigated in ref. [11] using discrete cosine transform (DCT) and principal component analysis (PCA) decorrelation methods. Some recent studies using the ASVspoof2015 corpus include pitch contour and strength of excitation [17] for spoofing detection.

In this work, our main goal was to demonstrate the effectiveness of Fisher vector representations derived from low-level descriptors (e.g., CQCC) of the signal for spoofing detection task. Fisher vector encoding was originally introduced in ref. [18] as a paradigm to build a discriminative classifier from a generative model. It was later identified as an effective utterance level representation technique for various computer vision applications such as image classification and large-scale image retrieval [19,20]. The basic idea behind Fisher vector encoding is to construct a generative model of local features (i.e., low-level descriptors) and use the gradient of the log-likelihood of a particular feature with respect to the model parameters as the feature's coding vector [18–20].

Fisher vector with cascaded non-linear normalization has been applied to classify the eating condition of a speaker from his/her recording for INTERSPEECH 2015 computational paralinguistic

challenge [21]. In this paper, we propose the use of Fisher vector encoding for voice spoofing detection task. Fisher vector can be used in combination with a linear support vector machine (SVM) or with a probabilistic linear discriminant analysis (PLDA) classifier. Here, we employ a PLDA backend for modeling bonafide and spoof classes.

The rest of the paper is organized as follows: Section 2 describes the low-level descriptor considered as countermeasures for this work. Sections 3 provides a description of the baseline spoofing detection systems considered herein. Fisher vector-based spoofing detection systems are presented in Section 4. Experimental results are reported and discussed in Section 5 and conclusions are drawn in Section 6.

2. Low-Level Acoustic Features as Spoofing Countermeasures

In the course of first (ASVspoof2015) and second (ASVspoof2017) speaker verification and countermeasure challenges and subsequently, it became clear that the most effective countermeasures against spoofing attacks are low-level acoustic feature (LLAF) or low-level descriptors (LLD). Low level descriptors (i.e., frame level features) are typically extracted at 10 ms intervals and designed to detect artifacts in spoof speech signal [4,11]. The most effective countermeasures for spoofing detection are the CQT-based cepstral coefficients (CQCC) [10], linear- frequency cepstral coefficients (LFCC), product spectrum-based cepstral coefficients (PSCC) [11–13], all pole group delay cepstral coefficients (APGDC) [13,22], linear prediction residual cepstral coefficients (LPRC) [13], and IIR-CQT-based cepstral coefficients (ICQC) [11]. In this work, we choose only CQCC low-level descriptors as countermeasures for spoofing detection on the ASVspoof2015 challenge data. Note that, low level descriptors are also known as local or frame level features.

Figure 2 presents a block diagram for the extraction of CQCC features as described in ref. [10]. After estimating CQT spectra, logarithmic compression is applied. A spline interpolation is applied to the estimated spectra to convert the geometric frequency scale to a linear scale. Finally, CQCC features are obtained by applying the discrete cosine transform. Like ref. [10], the number of bins per octave was set to $b = 96$ so that the corresponding quality factor is $Q = 1/\left(2.\hat{}\,(1/b) - 1\right) = 138$.

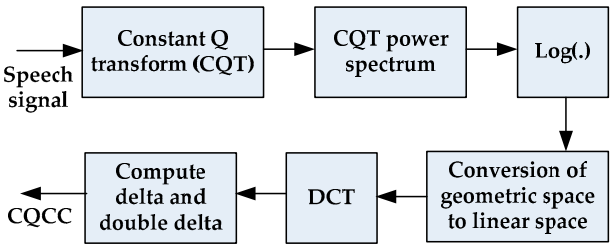

Figure 2. Extraction of constant Q cepstral coefficient (CQCC) features as proposed in ref. [10]. 40- dimensional delta + double delta coefficients are used as countermeasures.

3. Baseline Voice Spoofing Detection Systems

For the spoofing detection task, the de facto standard is the Gaussian mixture model (GMM) classifier trained using maximum likelihood training and low-level descriptor (e.g., CQCC) combination. To compare the performance of our proposed approach we build GMM- and i- vector/PLDA-based spoofing detection paradigms.

3.1. GMM-Based Framework

In a stand-alone GMM-based system (as shown in Figure 3), given the feature vector sequence O of a test speech signal, the bonafide versus spoofed speech decision is made based on the following log-likelihood ratio $l(O)$

$$l(O) = \log \frac{p(O|\lambda_b)}{p(O|\lambda_s)} = \log p(O|\lambda_b) - \log p(O|\lambda_s) = l_b - l_s \quad (1)$$

where λ_b and λ_s represent GMMs for bonafide and spoof classes, respectively; $l_b = \log p(O|\lambda_b)$ and $l_s = \log p(O|\lambda_s)$ are the average log-likelihood across all frames of the test speech signal obtained using bonafide and spoof models, respectively.

Based on CQCC features and with 32- and 512-Gaussian components GMMs (with diagonal covariance) we build two baseline systems and denote them as CQCC-GMM32 and CQCC-GMM512, respectively.

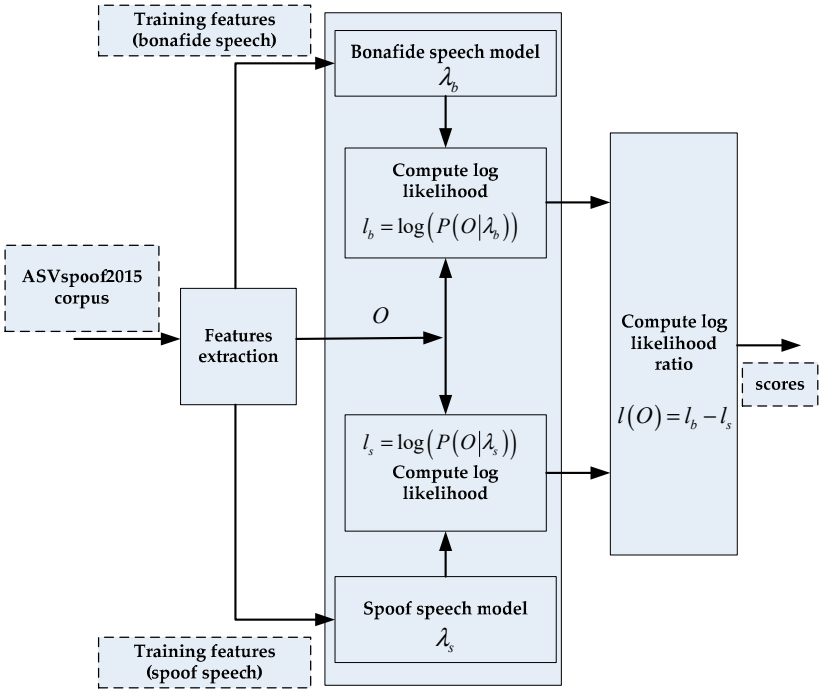

Figure 3. A block diagram showing various steps of the Gaussian mixture model (GMM)-based stand- alone spoofing detection paradigm.

3.2. I-vector/PLDA Framework

The conventional i-vector representation, proposed in ref. [23], represents the dominant approach in the speaker recognition field. Fundamentally, i-vector is the compact and fixed-length vector representation of a recording of arbitrary duration. In the i-vector/PLDA framework, for a chosen low-level descriptor, a 512-Gaussian components diagonal covariance universal background model (UBM) is trained on the ASVspoof2015 training data. After that, a 400-dimensional i-vector extractor is trained on the sufficient statistics generated from the training data. Binary classification can be then performed using a generative approach such as probabilistic linear discriminant analysis (PLDA) or a discriminative setting such as with support vector machines. This system is built on top of CQCC features with a PLDA backend and used in this work as baseline. Besides, we use the system CQCC-A from ref. [10] as baseline and denote it here as CQCC(A)-GMM512. This baseline system employs 20-dimensional CQCC features (acceleration coefficients only) as spoofing countermeasures.

4. Spoofing Detection using Fisher Vector Encoding

In Figures 4 and 5, we present schematic diagrams of the proposed voice anti-spoofing approaches based on Fisher vector encoding of the low-level descriptors. As local descriptors we use 40-dimensional CQCC features. In this section, we provide a description of various steps of the proposed methods.

This Fisher vector (FV) encoding is originally introduced and popularly used in computer vision, especially in large scale image retrieval [18–20]. The FV with cascaded non-linear normalization has also been applied to classify the eating condition of a speaker from his/her recording for INTERSPEECH 2015 computational paralinguistic challenge [21]. To the best of our knowledge, this encoding has not been applied for the voice anti-spoofing task.

The main idea behind this encoding is to measure the amount of change induced by the utterance/video descriptors on a background probability model, which is typically a Gaussian Mixture Model (GMM). Fisher vector encodes the amount of change of model parameters to optimally fit the new-coming data. This requires the computation of the Fisher information matrix, which is the derivative of the log likelihood with respect to model parameters (hence the name "Fisher"). This encoding requires a smaller number of components in a GMM than i-vector representation [21]. With Fisher vector, for modeling probabilistic linear discriminant analysis and support vector machine classifiers are used.

The FV encoding assumes that descriptors are generated by a GMM model with diagonal covariance matrices. At first, a *K*-Gaussians GMM model is learned on the training set. The GMM model is parameterized as $\lambda = \{w_k, \mu_k, \sigma_k\}_{k=1}^{K}$, where w_k, μ_k, and σ_k represent mixture weight, mean and variance corresponding to the *k*-th Gaussian component, respectively. Once the model is trained, the FV representation of a set of local descriptors $X = \{x_1, x_2, ..., x_N\}$ is given by the two parts [24]:

$$\mathbf{u}_k = \frac{1}{N\sqrt{w_k}} \sum_{i=1}^{N} q_{ki} \left(\frac{x_i - \mu_k}{\sigma_k} \right) \quad (2)$$

$$\mathbf{v}_k = \frac{1}{N\sqrt{2w_k}} \sum_{i=1}^{N} q_{ki} \left\{ \left(\frac{x_i - \mu_k}{\sigma_k} \right)^2 - 1 \right\} \quad (3)$$

where q_{ki} is the Gaussian soft assignment of the descriptor x_i to the *k*-th Gaussian. The **u** part captures the 1st order differences whereas the **v** part captures the 2nd order differences. With a *d*- dimensional local descriptor, the final representation of size $2dK$ is obtained by concatenation of the two parts. The dimension of extracted FV encoding is normally high. With K = 32 Gaussian components GMM and 40-dimensional local descriptors, the final dimension becomes 2*40*32 = 2560. So, we apply a principal component analysis (PCA) algorithm on the raw FV encoding to reduce the dimension to 600. Here, the PCA projection matrix is trained on the training data.

Power normalization followed by L_2-normalization are then applied [18–21]. Power normalization helps to reduce the sparsity of the descriptor and L_2-normalization aids in improving prediction performance. We utilize a component-wise power normalization with a = 0.5 as:

$$f(x) = sign(x)|x|^a \quad (4)$$

where $0 \le a \le 1$ is an optimization parameter.

In Figure 4, we present an overview of one of the voice anti-spoofing approach based on FV encoding of the low-level descriptors (e.g., CQCC). The approach is comprised of two sub-systems where, one sub-system is based on bonafide GMM (i.e., GMM trained using bonafide training data) and the other sub-system is based on spoof GMM (i.e., GMM trained using spoof training data). Binary classification is then performed using a generative approach such as probabilistic linear discriminant analysis (PLDA) backend. Final scores are obtained by using sum fusion of sub-systems' scores. Here, we denote this approach as FV1-PLDA.

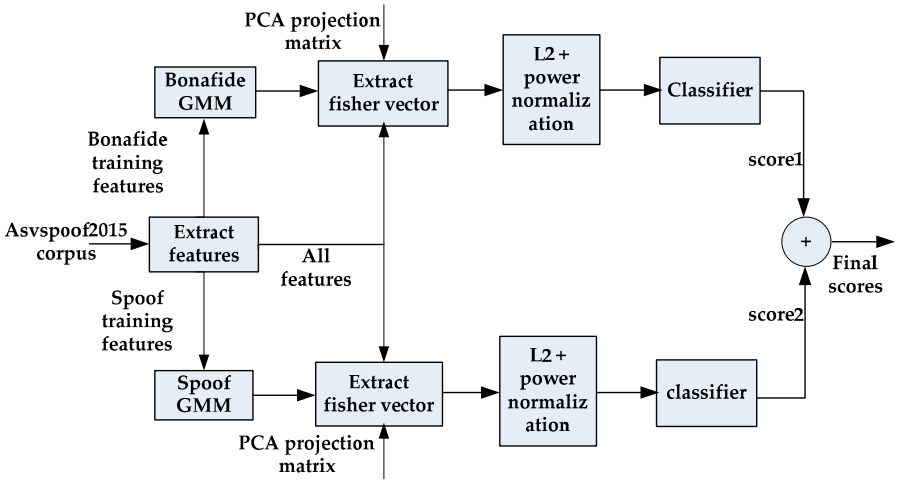

Figure 4. Schematic diagram showing various steps of Fisher vector encoding based spoofing detection systems. This approach is comprised of two sub-systems based on two Gaussian mixture models (GMMs) trained on bonafide and spoof training data, respectively. Here, principal component analysis (PCA) projection matrix is trained on the training.

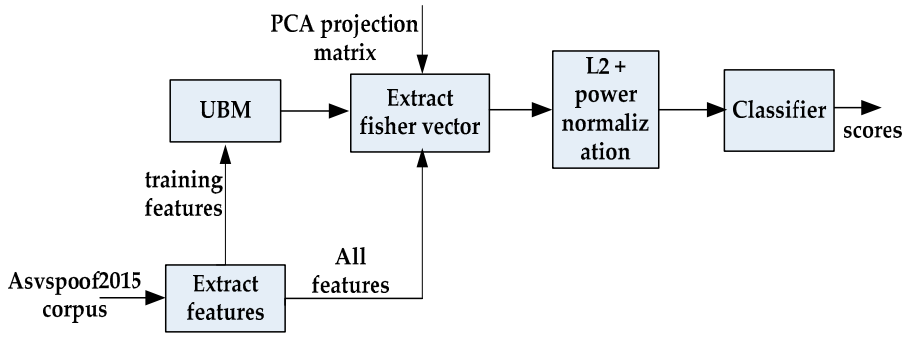

Figure 5. Schematic diagram showing various steps of Fisher vector encoding based spoofing detection systems. In this approach, a universal background model (UBM) is trained on the entire (bonafide + spoof) training data. Here, PCA projection matrix is trained on the training data.

In the second approach, as presented in Figure 5, instead of two GMMs, a single GMM (i.e., a universal background model (UBM)) is trained on the pooled bonafide + spoof training data. For modeling bonafide and spoof classes, a PLDA backend is used. This approach is denoted here as FV2- PLDA.

In both approaches, extraction of Fisher vectors and spoofing detection are carried out using the following steps:

1. Train GMMs on the training features (spoof and bonafide GMMs for the first approach and a UBM for the second approach)
2. Extract 2560-dimensional raw Fisher vectors (i.e., without L2 + power normalization) from the training data and train a PCA projection matrix on the extracted training Fisher vectors.
3. Extract 600-dimensional normalized Fisher vectors from all data using extracted local descriptors, trained GMMs and PCA projection matrix.
4. Perform binary classification using PLDA classifiers (or backend).
5. Perform score level fusion (for Figure 4 only).

Motivation behind Using Fisher Vector Encoding for Spoofing Detection

In the course of the first ASV spoof challenge (held in 2015) and subsequently it became clear that the dynamic coefficients (e.g., delta, acceleration or delta + acceleration) as countermeasure are more effective for logical access spoofing detection than the static and/or combination of static + dynamic coefficients. That is because spoofing techniques focus on modeling a smooth version (both temporal and spectral) of natural speech, which means a lack of temporal dynamic and missing spectral details. The Fisher Vector (FV), on the other-hand, encodes the gradients of the log-likelihood of the features under the Gaussian Mixture Model (GMM), with respect to the GMM parameters. This motivated us to employ FV encodings for the detection of logical access spoofing attacks.

5. Experiments and Results

In this section, we describe the ASVspoof2015 corpus, experimental setups, evaluation metric, experimental results on the evaluation set of ASVspoof2015 corpus together with discussion.

5.1. ASVspoof2015 Corpus

The ASVspoof2015 corpus is comprised of ten spoofing attacks denoted by S1, S2, …, S10. These spoofing attacks are mainly generated using various speech synthesis and voice conversion techniques. S1–S5 attacks are referred to as known and S6–S10 are called unknown attacks. There are three subsets in the data: training, development, and evaluation. The training set contains 16,375 recordings of which 3750 are bonafide (or genuine) and 12,625 are spoofed. In the development set there are 53,372 trials of which 3497 are bonafide and 49,875 spoofed trials and the evaluation set is comprised of 193,404 trials in total. Among them, 9404 trials are bonafide and 184,000 spoofed. We use the development set to tune the parameters of the systems and we report our results on the evaluation set. S3, S4, and S10 attacks are based on various speech synthesis algorithms and the rest of the attacks are generated using different voice conversion techniques. Among them, S8 spoofing attacks are based on Tensor based voice conversion [25] and S10 are Speech synthesis spoofing attacks generated using the MaryTTS toolkit [26]. For more detail about the corpus and type of spoofing attacks please see ref. [4].

5.2. Experimental Setup

In order to evaluate the performance of the proposed approach based on Fisher vector encoding, we carried out spoofing detection experiments on the ASVspoof2015 corpus [4] and the results are reported on the evaluation set. Our baseline systems are based on a standard GMM backend using the CQCC low-level descriptors. As an additional baseline, we also build an i-vector/PLDA spoofing detection paradigm on the top of CQCC features. With Fisher vector and i-vector representations, for modeling bonafide and spoof classes, a probabilistic linear discriminant analysis (PLDA) backend is used.

5.3. Evaluation Metrics

The equal error rate (EER) is used as a metric for performance assessment of the spoofing countermeasures. Spoofing detection scores are evaluated against each spoofing attack as well as using a single decision threshold (common to all spoofing attacks) and results are reported for ***Known*** (average over known spoofing attacks S1–S5), ***Unknown*** (average over unknown spoofing attacks S6–S10), ***Average*** (average over all ten spoofing attacks S1–S10) and ***All*** conditions. In ***All*** condition, the EERs are computed by evaluating spoofing detection scores against a single decision threshold (common to all spoofing attacks). This is because, in the real-applications scenario it is difficult to have any prior knowledge about the type of spoofing attacks.

5.4. Implementation

Local descriptors (i.e., CQCC features) and Fisher vector global descriptors are extracted in MATLAB and then converted to KALDI format. After that, GMM-based spoofing detection, extraction of i-vectors and binary classification using PLDA classier are conducted using the KALDI toolkit [24].

5.5. Results and Discussion

Fisher vector-based voice spoofing detection approaches i.e., FV1-PLDA and FV2-PLDA systems, are tested for a range of PCA (principal component analysis) dimensions and for a range of Gaussian components $C = \{8, 16, 32, 64, 128, 256, 512\}$ in a GMM. The best performance (i.e., lowest EER) on the development set is achieved with PCA dimension of 600 and 32-Gaussian components GMM. These optimal parameters are then used for reporting results on the evaluation set. For baseline GMM and i-vector/PLDA spoofing detection systems, the optimal number of Gaussian components is 512. But for a fair comparison with the Fisher vectors-based system we also report results of GMM baseline system with 32-Gaussian components. We also provide a comparison of performances when diagonal covariance versus full covariance GMMs are used in a stand-alone GMM-based spoofing detection framework. Below, we briefly summarize the spoofing detection systems considered in this work:

CQCC-GMM512: GMM-based spoofing detection system with 512-Gaussian components diagonal covariance GMMs on the top of 40-dimensional CQCC features.
CQCC-GMM512 (FC): Same as **CQCC-GMM512** but with full covariance GMMs.
CQCC-GMM32: Same as **CQCC-GMM512** but employs 32-Gaussian components GMMs.
CQCC-GMM32 (FC): Same as **CQCC-GMM512 (FC)** but uses 32-Gaussian components GMMs.
FV1-PLDA: Fisher vector-based spoofing detection system as presented in Figure 4.
FV1-PLDA (no norm): Same as **FV1-PLDA** but without using L_2 and power normalization methods.
FV2-PLDA: Fisher vector-based spoofing detection system employing a UBM as presented in Figure 5.
CQCC (A)-GMM512 [10]: This system is taken from ref. [10] and uses 20-dimensional CQCC features (acceleration coefficients only) as spoofing countermeasures with a 512-Gaussian component diagonal covariance GMM.
i-vector/PLDA: In this system, 400-dimensional i-vectors extracted with a 512-Gaussian component diagonal covariance UBM and CQCC features.
Fused: Equal weighted score level fusion (i.e., sum fusion) of **CQCC-GMM512, FV1-PLDA, FV2- PLDA** and **i-vector/PLDA** systems.

In Table 1, we present EERs attained by the GMM-based spoofing detection system with diagonal and full covariance GMMs. It is observed that the full covariance GMM-based system performs significantly better than the diagonal covariance-based system. In Table 2, we report EERs obtained using our proposed **FV1-PLDA** system when Fisher vectors are normalized using power normalization followed by L_2-normalization and when no-normalization is applied (denoted here as **FV1-PLDA (no norm)**). It is observed from Table 2 that L_2-normalization and power normalization helped to boost system performance significantly.

In Table 3, we report EER results achieved by all the spoofing detection systems including our proposed Fisher vector-based approaches (FV1-PLDA and FV2-PLDA). We can see from this table that the FV1-PLDA system outperforms the baseline CQCC-GMM32 and i-vector/PLDA systems. The best performance is demonstrated by the GMM-based system **CQCC-GMM512 (FC)** built with 512- Gaussian component full covariance GMMs. **Fused** system outperformed all other systems based on diagonal covariance GMM or UBM.

Table 1. Comparison of spoofing detection performance (in terms of equal error rates (EER)) of Gaussian mixture model (GMM)-based systems when diagonal covariance and full covariance GMMs are used. The best results are highlighted in bold face.

	Known	Unknown	Average	All
CQCC-GMM512	0.2375	0.796	0.517	0.836
CQCC-GMM512 (FC)	**0.034**	**0.365**	**0.199**	**0.433**

Table 2. Comparison of spoofing detection performance (in terms of EER) of the Fisher vector-based systems with power normalization followed by L_2 normalization (FV1-PLDA) and without any normalization (FV1-PLDA (no norm)). The best results are highlighted in bold face.

	Known	Unknown	Average	All
FV1-PLDA	**0.114**	**1.223**	**0.668**	**1.048**
FV1-PLDA (no norm)	0.228	5.697	2.963	3.935

Table 3. Spoofing detection performance (in terms of EER) of all systems considered herein, including the Fisher vectors-based systems on the ASVspoof2015 evaluation set. Results are reported on the *Known* (average of S1–S5), *Unknown* (average of S6–S10), *Average* (average of S1–S10), and *All* conditions (using a single decision threshold) as described in Section 5.3. The lowest EERs are highlighted in bold face. In [10], the performance of **CQCC (A)-GMM512** was not reported on the *All* evaluation condition, and therefore, we were not able to compare it in this work.

	Known	Unknown	Average	All
CQCC-GMM32	0.489	1.138	0.814	1.084
CQCC-GMM32 (FC)	0.164	0.669	0.417	0.618
FV1-PLDA	0.114	1.223	0.668	1.048
FV2-PLDA	0.597	3.664	2.131	2.951
CQCC-GMM512	0.2375	0.796	0.517	0.836
CQCC (A)-GMM512 [10]	0.048	0.462	0.255	-
CQCC-GMM512 (FC)	**0.034**	**0.365**	**0.199**	**0.434**
i-vector/PLDA	0.161	4.017	2.089	2.826
Fused	**0.0243**	**0.307**	**0.165**	**0.456**

These results are motivating, and it shows the benefit of using Fisher vector encoding and full covariance GMMs instead of diagonal covariance GMMs for the voice anti-spoofing task. The EERs obtained using all systems considered herein, are reported in Tables 4 and 5 for each of the known (S1–S5) and unknown (S6–S10) spoofing attacks, respectively. Note that, spoof signals of S10 attack were produced by concatenating the selected units from the time-domain signal without any vocoding technique. The major artifacts of S10 are due to the discontinuities introduced at the joint of two selected units [4,10–11]. Since no vocoder was used in the S10 spoofing technique synthesis [4], vocoder mismatch between the training and evaluation data resulted in significantly higher EERs for all spoofing countermeasures on the S10 attack compared to the other nine (S1–S9) attacks.

This was also the case in the ASVspoof2015 challenge where the EERs for all participants were high (greater than 8%) for S10. The main reason behind such poor performance was the use of VAD and auditory filterbank in the feature extraction framework [11]. Removal of non-speech frames causes some spoofing artifacts to be removed, specifically for unknown spoofing attacks (S6–S10), while integration of auditory filterbank causes some of the artifacts present in the raw spectra to be smoothed out which resulted in an increase the error rates [10,11]. It can be seen from Figures 4 and 5 that, all systems performed better on known attacks than on unknown attacks. By looking at the EERs on S10 attack in Table 5, we can see that Fisher vectors and i-vector-based systems performed worse on S10 attack. One interesting conclusion we can draw from this observation is that global descriptors (i.e., utterance level embedding) extracted from the local descriptors (e.g., CQCC) cause smoothing out or elimination of artifacts present in the local descriptors.

Table 4. Spoofing detection performance (in terms of EER) of all systems considered herein, including the Fisher vectors-based systems on the ASVspoof2015 evaluation set. Scores are evaluated against each spoofing attack (S1–S10) and results are reported on each of the ***known*** attacks. The lowest EERs are highlighted in bold face.

	Known Attacks				
	S1	**S2**	**S3**	**S4**	**S5**
CQCC-GMM32	0.124	1.203	0.043	0.0465	1.0286
CQCC-GMM32 (FC)	0.047	0.373	0.000	0.004	0.396
FV1-PLDA	0.069	0.253	0.019	0.017	0.208
FV2-PLDA	0.664	1.036	0.276	0.300	0.709
CQCC-GMM512	0.062	0.609	0.008	0.0184	0.489
CQCC (A)-GMM512 [10]	**0.005**	0.106	**0.000**	**0.000**	0.130
CQCC-GMM512 (FC)	**0.010**	0.064	0.011	0.011	**0.074**
i-vector/PLDA	0.141	0.318	0.053	0.047	0.244
Fused	**0.015**	**0.069**	**0.000**	**0.000**	**0.037**

Table 5. Spoofing detection performance (in terms of EER) of all systems considered herein, including the Fisher vectors-based systems on the ASVspoof2015 evaluation set. Scores are evaluated against each spoofing attack (S1–S10) and results are reported on each of the ***unknown*** attacks. The lowest EERs are highlighted in bold face.

	Unknown attacks				
	S6	**S7**	**S8**	**S9**	**S10**
CQCC-GMM32	1.011	0.366	2.495	0.584	1.237
CQCC-GMM32 (FC)	0.313	0.254	1.436	0.280	1.062
FV1-PLDA	0.244	0.064	1.892	0.151	3.760
FV2-PLDA	0.995	0.389	2.242	0.458	14.236
CQCC-GMM512	0.455	0.227	2.015	0.323	**0.961**
CQCC (A)-GMM512 [10]	**0.098**	**0.064**	**1.033**	**0.053**	1.065
CQCC-GMM512 (FC)	**0.089**	**0.046**	**1.165**	**0.0558**	**0.469**
i-vector/PLDA	0.331	0.285	12.345	0.329	6.795
Fused	**0.057**	0.027	0.692	0.037	**0.720**

6. Conclusions

In this work, we proposed the use of Fisher vector encoding of low-level descriptors, such as constant Q cepstral coefficients, as countermeasure for the spoofing detection task. We also investigated the performance of diagonal and full covariance GMMs in a stand-along GMM-based spoofing detection system. In order to evaluate performance, we carried out standalone spoofing detection experiments and reported results on the evaluation set of the ASVspoof2015 challenge corpus. Our proposed system outperformed the i-vector/PLDA system and GMM systems built with 32-Gaussian component diagonal covariance GMMs. The Fisher vector-based system is simple and computationally efficient as it requires a smaller number of Gaussian components in a GMM. The GMM-based system with a large number of Gaussian components (e.g., 512) and full covariance GMMs performed the best. Score level fusion of the selected four sub-systems helped to further reduce EER.

Acknowledgments: The authors wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2019-05381. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the NSERC.

References

1. Evans, N.; Kinnunen, T.; Yamagishi, J.; Wu, Z.; Alegre, F.; de Leon, P. Speaker recognition anti-spoofing. In *Handbook of Biometric Anti-Spoofing*; Marcel, S., Li, S., Nixon, M., Eds.; Springer-Verlag: London, UK, 2014.
2. Kinnunen, T.; Wu, Z.; Lee, K.A.; Sedlak, F.; Chng, E.S.; Li, H. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4401–4404.
3. de Leon, P.L.; Pucher, M.; Yamagishi, J. Evaluation of the vulnerability of speaker verification to synthetic speech. In Proceedings of the IEEE Speaker and Language Recognition Workshop (Odyssey), Brno, Czech Republic, 28 June–1 July 2010; pp. 151–158.
4. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The First ASV Spoofing and Countermeasures Challenge. In Proceedings of the INTERSPEECH 2015, Dresden, Germany, 6–10 September 2015. Available online: http://www.spoofingchallenge.org/is2015_asvspoof.pdf (accessed on 01 July 2019).
5. Chen, N.; Qian, Y.; Dinkel, H.; Chen, B.; Yu, K. Robust Deep Feature for Spoofing Detection—The SJTU System for ASVspoof 2015 Challenge. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015.
6. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
7. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 4006–4010.
8. Tamamori, A.; Hayashi, T.; Kobayashi, K.; Takeda, K.; Toda, T. Speaker-dependent WaveNet vocoder. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1118–1122.
9. Kaneko, T.; Kameoka, H.; Hojo, N.; Ijima, Y.; Hiramatsu, K.; Kashino, K. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In Proceedings of the 2017 ICASSP, New Orleans, LA, USA, 5–9 March 2017; pp. 4910–4914.
10. Todisco, M.; Delgado, H.; Evans, N. Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification. *Comput. Speech Lang.* **2017**, *45*, 516–535.
11. Alam, J.; Kenny, P. Spoofing Detection Employing Infinite Impulse Response—Constant Q Transform-based Feature Representations. In Proceedings of the EUSIPCO, Kos Island, Greece, 28 August–2 September 2017.
12. Alam, J.; Kenny, P.; Gupta, V.; Stafylakis, T. Spoofing Detection on the ASVSpoof2015 Challenge Corpus Employing Deep Neural Networks. In Proceedings of the Odyssey Speaker and Language Recognition Workshop, Bilbao, Spain, 21–24 June 2016.
13. Alam, J.; Kenny, P.; Bhattacharya, G.; Stafylakis, T. Development of CRIM System for the Automatic Speaker Verification Spoofing and Countermeasures Challenge 2015. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015. Available online: https://www.asvspoof.org/asvspoof2015/CRIM.pdf (accessed on 01 July 2019).
14. Patel, T.B.; Patil, H.A. Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015.
15. Xiao, X.; Tian, X.; Du, S.; Xu, H.; Chng, E.S.; Li, H. Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: The NTU Approach for ASVspoof 2015 Challenge. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015.
16. Tian, X.; Wu, Z.; Xiao, X.; Chng, E.S.; Li, H. Spoofing detection from a feature representation perspective. In Proceedings of the ICASSP, Shanghai, China, 20–25 March 2016.
17. Patel, T.; Patil, H. Effectiveness of fundamental frequency (F0) and strength of excitation (SOE) for spoofed speech detection. In Proceedings of the ICASSP, Shanghai, China, 20–25 March 2016.
18. Jaakkola, T.; Haussler, D. Exploiting generative models in discriminative classifiers. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 30 November–5 December 1998; pp. 487–493.

19. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
20. Perronnin, F.; Liu, Y.; Sanchez, J.; Poirier, H. Large-scale Image Retrieval with Compressed Fisher Vectors. In Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3384–3391.
21. Kaya; H; Karpov, A.; Salah, A.A. Fisher Vectors with Cascaded Normalization for Paralinguistic Analysis. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015.
22. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A Comparison of Features for Synthetic Speech Detection. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015.
23. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 788–798.
24. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society, Big Island, HI, USA, 11–15 December 2011.
25. Saito, D.; Yamamoto, K.; Minematsu, N.; Hirose, K. One-to-many voice conversion based on tensor representation of speaker space. In Proceedings of the Interspeech 2011, Florence, Italy, 27–31 August 2011; pp. 653–656.
26. The MARY TTS—An Open-Source, Multilingual Text-To-Speech Synthesis System. Available online: http://mary.dfki.de (the first version was released on 14 February 2006).