Features and Classifiers for Replay Spoofing Attack Detection

Cemal Hanilçi

Department of Electrical and Electronic Engineering, Bursa Technical University, Bursa, Turkey
cemal.hanilci@btu.edu.tr

Abstract

Automatic speaker verification (ASV) systems are known to be highly vulnerable against spoofing attacks. Various successful countermeasures have recently been proposed to detect spoofing attacks originating from speech synthesis (SS) and voice conversion (VC). However, detecting replay attacks, the most easily implementable spoofing attacks against ASV systems, has gained less attention. Thus, in this paper we present an experimental comparison of various feature extraction techniques and classifiers for replay attack detection. In total, six magnitude spectrum and three phase spectrum based features are used for feature extraction. For classification in turn, four different techniques are utilized. Experiments are conducted on recently released ASVspoof 2017 replay attack detection challenge. Experimental results reveals that magnitude spectrum features considerably outperform phase based features independent of the classifier. Comparative results using four different classifiers indicate that i-vector cosine scoring yields lower equal error rates (EERs) than other methods.

1. Introduction

Automatic speaker verification (ASV) is the task of automatically accepting or rejecting an identity claim given a speech signal [1]. Recent developments on ASV technology has led to an increasing range of applications where ASV systems are primarily used. However, it has independently been shown and confirmed that ASV systems are highly vulnerable against spoofing attacks [2, 3]. For any biometric person authentication system, spoofing attack refers to an attack where a fraudster masquarades herself in order to gain an illegitimate access to the system [4]. There exists four main types of spoofing attacks against ASV systems [5]: (i) *mimicricy* [6], (ii), *speech synthesis* (SS) [7], (iii) *voice conversion* (VC) [8] and (iv) *replay* [9]. Among these four types of attacks, mimicricy is less likely to occur since it requires a professional/talented mimicker. Speech synthesis is the task of synthesizing a target speaker's voice given a text input whereas voice conversion aims at modifying a source speaker's voice towards target speaker's voice. Replay in turn, is the attack where target speaker's prerecorded voice sample is used for accessing to the sytem.

Spoofing countermeasures – determining whether a speech signal is genuine or spoofed– have recently gained great interest in order to protect ASV systems against spoofing attacks. The studies related to anti-spoofing mostly focus on detecting the spoofed signals generated using SS or VC attacks due to *the Automatic Speaker Verification Spoofing and Countermeasures Challenge* organized as a special session in INTERSPEECH 2015 conference [10]. During the challenge, a database consisting of genuine and synthetic utterances were shared with the participants and they were asked to design countermeasures discriminating genuine speech from spoofed speech generated using various SS and VC techniques. Various countermeasures were proposed for synthetic speech detection and in general phase features with Gaussian mixture model (GMM) classifier were found to yield promising results [11, 12].

Countermeasures for detecting replay attacks in turn, has been less studied in comparison to SS and VC attacks. In contrast to other spoofing attack types (mimicricy, SS and VC), replay attacks are the easiest to implement and do not require any specific expertise or equipment. Therefore, they are most likely to occur in practice. The vulnerability of ASV systems utilizing various speaker modeling techniques against replay attacks have been studied in detail in [13]. In [14, 15], spectral ratio statistics (e.g. spectral ratio, low-frequency ratio, modulation index) were proposed as features for detecting the replay attacks from far field recordings and it was found that low-frequency ratio features yield perfect detection accuracy in matched channel case. However, in case of mismatched channel, the 7.32% equal error rate (EER) was reported for replay attack detection. In the same study, it was shown that frequency response of a standard loudspeaker attenuates the low frequency region considerably. Therefore low-frequency ratio features were found to be useful for replay attack detection. However, most of the studies have generally reported their findings using a small number of recording and playback conditions [14, 15, 9]. That was one of the motivations behind the organizing *ASVspoof 2017 Automatic Speaker Verification Spoofing and Cuntermeasures Challenge*[1].

In this paper, we focus on replay attack detection using ASVspoof 2017 database from both feature extraction and classification perspectives. To this end, we explore the performances of magnitude and phase features on replay attack detection using four different and well-known classifiers. Six magnitude spectrum based features and three phase based features are used in the experiments. The features are selected according to their performances on synthetic and converted speech detection as reported in [11, 16]. Thus in this study we aim at investigating the generalization capability of the countermeasures for both replay and SS/VC attacks. Though simple GMM classifier trained using maximum likelihood criterion was found to perform better for detecting SS and VC attacks [12], the performance of different classifiers remain unknown for replay attack detection. Therefore, in this study we compare four different classifiers on ASVspoof 2017 database.

This work was supported by the Bursa Technical University under project no. 2016-01-012

[1] `http://www.spoofingchallenge.org`

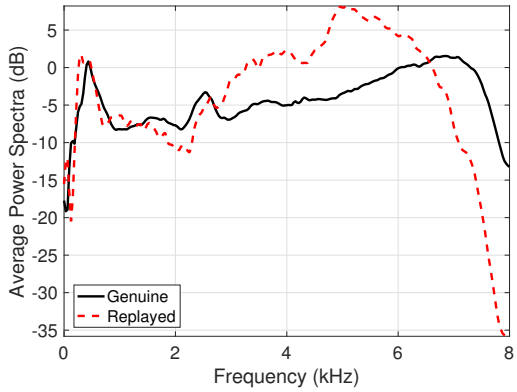

Figure 1. Average long term spectra for genuine and replayed speech signals of training set of ASVspoof 2017 database.

2. Feature Extraction Techniques

2.1. Magnitude Spectrum Features

In order to extract features, a speech signal is first divided into short overlapping frames (20 ms frames with 10 ms overlap used here). Then each frame is windowed with Hamming window and discrete Fourier transform (DFT) of each windowed frame is computed. The DFT of a windowed frame $x[n]$ can be represented as:

$$X(\omega) = |X(w)|\, e^{j\theta(\omega)} \tag{1}$$

where $|X(\omega)|$ and $\theta(\omega)$ are the magnitude and phase spectra of the $x[n]$ at frequency bin ω, respectively.

Figure 1 shows the average long term power spectra computed over 1508 genuine and 1508 replayed speech utterances using the training set of the ASVspoof 2017 database. From the figure, it can be observed that at genuine and replayed signals show similar behaviour at low frequency region. However, at high frequency region they show different characteristics. Therefore, intuitively magnitude spectrum based features would be useful for replay attack detection. We briefly explain the six different magnitude spectrum based features in the following.

2.1.1. Cepstrum

Simple cepstral coefficients (CEPs) are computed by applying discrete cosine transform (DCT) to the logarithm of the power spectrum, $|X(\omega)|^2$. Previously they were used for synthetic speech detection and found to give encouraging results [11].

2.1.2. Filterbank Cepstral Coefficients

We consider four types of filterbank based cepstral coefficients extracted from the DFT power spectrum, $|X(\omega)|^2$. First, the Mel-frequency cepstral coefficients (MFCCs) are extracted by processing the power spectrum through a 30-channel triangular filterbank spaced in Mel-scale. The MFCCs are obtained by applying DCT to the logarithmic filterbank outputs. The rectangular filter cepstral coefficients (RFCCs) are extracted similar to MFCCs. However, in RFCCs, the filterbank consisting of rectangular filters spaced in linear scale. Linear frequency cepstral coefficients (LFCCs) in turn, are computed the same way but the triangular filters are used rather than rectangular filters. Finally, for the inverted Mel-frequency cepstral coefficients (IMFCCs) [17], triangular filters are spaced linearly on the inverted Mel scale which puts higher emphasis to the higher frequency region.

2.1.3. Constant Q Transform Cepstral Coefficients

Constant Q cepstral coefficients (CQCCs) are another magnitude spectrum based features and they have recently been used for synthetic speech detection with encouraging results on ASVspoof 2015 database [18, 16]. Besides, it has been provided by the ASVspoof 2017 organizers as the baseline countermeasure [19]. In contrast to other filterbank cepstral coefficients where features are extracted from DFT spectra, CQCCs are extracted from wavelet-like perceptually motivated time-frequency analysis known as constant Q transform (CQT) [20]. The features are extracted from the uniformly sampled CQT power spectrum by following discrete cosine transform (DCT) [18, 16].

2.2. Phase Based Features

2.2.1. Cosine Phase Features

The phase spectrum, $\theta(\omega)$, computed from the DFT of the speech frames is first unwrapped because of the discontinuouty. Then the cosine function is applied to to the unwrapped phase which yields a normalized phase spectrum within the range [-1.0, 1.0]. Then DCT is applied to the normalized phase to obtain cosine phase features. This feature is used in spoofing detection in [21] and called as CosPhase features.

2.2.2. Modified Group Delay Function

Modified group delay function (MGDF) is computed from the speech frame $x[n]$ and defined as:

$$\tau(\omega) = \text{sgn} \times \left| \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}} \right|^{\alpha} \tag{2}$$

where $X_R(\omega)$ and $X_I(\omega)$ are the real and imaginary parts of the DFT of the windowed speech frame, $x[n]$, at frequency bin ω. $Y_R(\omega)$ and $Y_I(\omega)$ are the real and imaginary parts of the DFT of $nx[n]$. $S(\omega)$ is the smoothed version of the magnitude spectrum $|X(\omega)|$. α and γ are the control parameters and they are both set to 0.1. Features are extracted by applying DCT to the MGDF. The so-called MODGD features are used in synthetic speech detection in [21].

2.2.3. Relative Phase Shift Features

The relative phase shift (RPS) features [22] use the harmonic modeling of the speech signal where each frame is represented as the sum of sinusoids, $x[n] = \sum_k A[k]cos(\phi_k[n])$. Here, $A[k]$ is the amplitude and $\phi_k[n] = 2\pi kF_0n + \theta_k$ is the instantaneous phase of the kth harmonic. F_0 is the fundamental frequency and θ_k is the initial phase. The RPS value is defined as the phase shift of the kth harmonic with respect to fundamental frequency. The RPS features are extracted from the RPS values by first unwrapping the phase and then differentiating. The differentiated phase values are integrated by Mel-filterbank and they converted into features by applying DCT. RPS features were found to give encouraging results on synthetic speech detection [22, 11].

3. Replay Attack Detection

3.1. Gaussian Mixture Models

Gaussian mixture models (GMM) [23] is a popular and well-known classification technique mostly used in speech processing. In GMM, each class is represented as a weighted sum of M multi-variate Gaussians, $p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} w_i p_i(\mathbf{x})$. Here, $w_i,\ i = 1, \ldots, M$ are the mixture weights and $p_i(\mathbf{x})$ is the $D-$variate Gaussian density function with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$.

The parameters $(w_i,\ \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i)$ of a GMM are estimated by expectation maximization algorithm with maximum likelihood criterion. For replay attack detection, a GMM is trained for each class using the training data of the corresponding class. During the test phase, given the GMMs λ_{genuine} and λ_{replay}, and the feature vectors extracted from the test speech signal, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{\text{T}}\}$, the likelihood ratio detection score is computed as,

$$\Lambda(\mathbf{Y}) = L(\mathbf{Y}|\lambda_{\text{genuine}}) - L(\mathbf{Y}|\lambda_{\text{replay}}) \tag{3}$$

where, $L(\mathbf{Y}|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{y}_{\text{t}}|\lambda)$ is the average logarithmic likelihood of feature vectors $\mathbf{Y}$ given the GMM λ.

GMM is designated as the baseline classifier by the organizers of the ASVspoof 2017 challenge.

3.2. Support Vector Machines

Support vector machines (SVM) is a well-known binary classifier successfully used in speaker and language recognition. It tries to model the decision boundary between two classes as a separating hyperplane by maximizing the margin. From the replay attack detection point of view using SVM classifier, one class corresponds to high-dimensional supervectors of genuine class and other class corresponds to the supervectors of replay class which are labeled as $+1$ and -1, respectively.

Two different supervector extraction methods are used for the replay attack detection with SVM classifier: *generalized linear discriminant sequence kernel* (GLDS-SVM) [24] and *local binary patterns* (LBP) [25]. In GLDS, the feature vectors are mapped into a high-dimensional space by polynomial expansion.

The LBP approach was first introduced for texture classification with promising results [25]. It is defined as the binary comparisons of the pixel intensities between the center pixel and its eight neighbourhood. For speech applications, LBP operator is applied to the feature matrix consisting of feature vectors extracted from each frame rather than image [26]. The resulting matrix is called *textrogram*. The LBP supervector is computed by first concatenating the histograms of the pixel values within the each row in the textrogram. The histograms are then normalised and their corresponding bin values are stacked vertically to obtain the supervector.

3.3. i-vector Approach

i-vector approach has become state-of-the-art for speaker recognition in the past years [27]. It extracts a low-dimensional vector, $\mathbf{w}$ referred to an *i-vector* from a speech signal by factorizing the GMM mean supervector $\boldsymbol{\mu}$ as $\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{w}$ [27]. Here, $\mathbf{m}$ is the mean supervector comes from the universal background model (UBM), $\mathbf{T}$ is a low-rank variability matrix representing the total variability subspace and $\mathbf{w}$ is the low-dimensional i-vector with standard normal prior distribution.

For replay attack detection, the i-vectors of training utterances for both genuine and replay classes are extracted. The extracted i-vectors are pre-processed by applying within class covariance normalization (WCCN) [28]. For replay attack detection we aim at removing or normalizing the within-class (genuine or replay) variation. Therefore WCCN transformation matrix is computed from the training i-vectors of both classes (genuine and replay). Normalized i-vectors are then projected onto unit-sphere by applying length normalization [29].

Since there are multiple training i-vectors for each class in ASVspoof 2017 database, we represent genuine and replay classes with their average training i-vectors as $\mathbf{w}_{\text{genuine}} = (1/K)\sum_{k=1}^{K} \mathbf{w}_{\text{genuine}}^{k}$, where $\mathbf{w}_{\text{genuine}}^{k}$ is the kth training i-vector of genuine class and K is the total number of training utterances. Similarly, average training i-vector, $\mathbf{w}_{\text{replay}}$ is computed for replay class.

In the detection step, given the i-vector extracted from the test utterance, $\mathbf{w}_{\text{test}}$ we use *cosine similarity* as the detection score and it is computed as:

$$\text{score} = \mathbf{w}_{\text{genuine}}^{\text{T}}\mathbf{w}_{\text{test}} - \mathbf{w}_{\text{replay}}^{\text{T}}\mathbf{w}_{\text{test}}. \tag{4}$$

Since $\|\mathbf{w}_{\text{genuine}}\| = \|\mathbf{w}_{\text{replay}}\| = \|\mathbf{w}_{\text{test}}\| = 1$ because of the length normalization, it is omitted in (4).

4. Experimental Setup

4.1. Database

The ASVspoof 2017 database is used in the experiments [19]. The database consists of three disjoint subsets: *Training*, *Development* and *Evaluation*. Training set includes 1508 genuine and 1508 spoofed (replayed) speech signals from 10 male speakers where replayed signals are generated with three different replay configuration. The training set is used to train classifiers and to estimate the hyperparameters of the countermeasures. *Development* set consists of 760 genuine and 950 spoofed speech signals from a total of 8 speakers. The replay devices used to generate development set are mostly different from those used to generate training set. The development set is used to optimise the countermeasures. *Evaluation* set in turn, includes 1298 genuine and 12922 replayed utterances from 42 speakers. More details about the database and recording condition can be found in Evaluation Plan of the ASVspoof 2017 challenge in [19].

4.2. Performance Criterion

Equal error rate (EER) is the primary performance criterion for ASVspoof 2017 challenge [19]. EER corresponds to the error rate for the threshold where false alarm (P_{fa}) and miss rate (P_{miss}) are equal. P_{fa} is the ratio of the number of replayed trials determined as genuine to the total number of replayed trials. Similarly, P_{miss} is the ratio of the number of genuine trials determined as replayed to the total number of genuine trials. The reported EERs are computed using the Bosaris toolkit[2] which uses the receiver operating characteristics convex hull (ROCHH) to estimate the EER.

4.3. Features and Classifiers

In the experiments, nine different features (six magnitude spectrum and three phase based) described in Section 2 are

[2] `https://sites.google.com/site/bosaristoolkit/`

used. 19 features (including c_0) and their first and second order derivatives (dynamic coefficients) which yields a total of 57 dimensional features are used for all feature extraction methods. This configuration is selected based on the initial experiments. Similarly, according to initial experiments on development set, it was found that applying voice activity detection (VAD) reduces the replay attack detection performance considerably. Therefore, VAD is not applied in the experiments.

In the experiments, we use four different classification methods: GMM, GLDS-SVM, LBP-SVM and i-vector. In GMM method, for each class (genuine and replay) a GMM with diagonal covariance matrices consisting of 512 Gaussian components are trained with 10 EM iterations. For GLDS-SVM, the polynomial expansion of order $m = 3$ is used. For LBP-SVM, 8-1 configuration, binary comparison of center pixel with its 8 neighbourhood, is used. For both GLDS and LBP, linear kernel SVM is trained using LIBSVM package [30]. Universal background model (UBM) with 512 Gaussians is trained using the entire training set of the ASVspoof 2017 database for the i-vector system. The same training data is used to train the i-vector extractor (T-matrix) and 600 dimensional i-vectors are extracted in the experiments.

5. Results

The experiments are first conducted on development set. The EERs obtained with different features and classifiers on Development set are summarized in Table 1. From the results, we find that in contrast to synthetic speech detection [11], magnitude features considerably outperform phase based features in replay attack detection independent of the classifier. Among four filterbank features, MFCC yields the highest EERs. Another interesting observation is that simple cepstral features yields promising results and in some cases they are superior to other features. For example they give approximately 27% and 45% better performance than MFCC and CQCC features with GMM classifier, respectively. For phase based features, MODGD is superior to others on replay attack detection. Interestingly, the highest EERs are obtained with RPS features. This is possibly because of the fact that RPS is calculated as the phase shift of the harmonics with respect to fundamental frequency. However, replay introduces convolutive distortion on the original speech signal. Thus estimating the fundamental frequency from convolutionally distorted speech becomes difficult.

From the classifier perspective, it can be seen that i-vector modelling yields smaller EERs than other classifiers for the majority of the nine features (except for IMFCC). For IMFCC features, GMM shows better replay attack detection performance than SVM and i-vector classifiers. The smallest EER of 4.56% is achieved with RFCC features using i-vector system, whereas IMFCC features with GMM classifier gives the EER of 4.71%.

The last row of the Table 1, shows the EERs obtained by applying score fusion to the all features for each classifier. It can be observed that score fusion improves the performance for GMM and i-vector systems. For GMM classifier, score fusion yields approximately 15% performance improvement over the best performing countermeasure (EER reduced from 4.71% to 4.01%). Whereas approximately 29% (EER $4.56\% \rightarrow 3.24\%$) relative performance improvement is observed after appyling score fusion to the i-vector countermeasures. Interestingly, for GLDS and LBP SVM systems, score fusion does not bring any improvements on the best performing systems.

Next, experiments are carried out on the Evaluation set of the ASVspoof 2017 database using GMM and i-vector systems. The EERs for each individual features and their fusion are given in Table 2. Similar to observations on development set, i-vector system is superior to GMM classifier on Evaluation set except for RPS features. CQCC features with i-vector system achieves the smallest EER of 21.38% on the evaluation set. In contrast to findings on development set, score fusion does not bring any performance improvement over the best performing system (IMFCC for GMM and CQCC for i-vector). This is possibly because the fusion weights were trained using the development set scores. However, evaluation set includes spoofed utterances generated using different replay configurations than the ones used to generate development set.

Table 1. EERs (%) for different features and classifiers on development set.

	Classifier			
Features	GMM	GLDS-SVM	LBP-SVM	i-vector
MFCC	10.34	13.66	18.98	**9.71**
LFCC	6.08	9.85	7.37	**4.58**
RFCC	6.91	6.57	7.93	**4.56**
IMFCC	**4.71**	12.02	17.43	6.73
CQCC	11.85	9.43	9.03	**8.85**
CEPS	8.14	11.57	8.55	**4.60**
MODGD	11.19	14.18	9.03	**4.78**
COSPHASE	25.95	23.99	17.10	**11.90**
RPS	43.36	40.06	**39.91**	41.26
Fusion	4.01	11.37	10.67	**3.24**

Table 2. The results on evaluation set.

	Classifier	
Features	GMM	i-vector
MFCC	33.14	**29.42**
LFCC	33.18	**27.00**
RFCC	32.42	**30.03**
IMFCC	30.87	**23.44**
CQCC	32.27	**21.38**
CEPS	36.24	**28.48**
MODGD	36.65	**26.95**
COSPHASE	46.90	**36.00**
RPS	**26.65**	32.47
Fusion	30.51	**24.20**

6. Conclusions

In this paper, an extensive study with different feature extraction and classification techniques have been performed for replay attack detection to protect ASV systems. Experimental results indicate that magnitude features are more useful than the phase features for detecting replay attacks in general. The RFCC features which uses rectangular filters yields the best performance on development set with i-vector system. However, CQCC features were found to give the smallest EER in comparison to other eight different feature extraction methods on evaluation set. It was observed that score fusion improves the replay attack detection on development set but it does not bring any performance improvement on evaluation set.

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Proc. BTAS*, 2015, pp. 1–6.

[3] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proc. APSIPA ASC*, 2012, pp. 1–5.

[4] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

[5] 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.

[6] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[7] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

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

[10] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[11] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. INTERSPEECH*, 2015, 2087–2091.

[12] C. Hanilçi, T. Kinnunen, and M. Sahidullah, "Classifiers for synthetic speech detection: A comparison," in *Proc. INTERSPEECH*, 2015, pp. 2057–2061.

[13] A. Janicki, F. Alegre, and N. Evans, "An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Security and Communication Networks, Wiley, 2016*, pp. 3030–3044, 06 2016.

[14] J. Villalba and E. Lleida, *Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 274–285.

[15] ——, "Preventing replay attacks on speaker verification systems," in *2011 Carnahan Conference on Security Technology*, Oct. 2011, pp. 1–8.

[16] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, pp. –, 2017.

[17] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing*, vol. 4, no. 2, pp. 114–122, 2007.

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

[19] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," http://www.spoofingchallenge.org/data2017/asvspoof_2017_evalplan_v0.pdf, online; accessed 13th March 2017.

[20] J. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

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

[22] J. Sánchez, I. Saratxaga, I. Hernáez, 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.

[23] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[24] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[25] T. Ojala, M. PietikÃinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.

[26] F. Alegre, R. Vipperla, A. Amehraye, and N. W. D. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 940–944.

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

[28] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. ICSLP*, 2006.

[29] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252.

[30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.