International Journal of Speech Technology (2019) 22:271–281
https://doi.org/10.1007/s10772-019-09602-z

# Replay spoofing countermeasures using high spectro-temporal resolution features

**K. N. R. K. Raju Alluri[1] · Anil Kumar Vuppala[1]**

Received: 6 November 2018 / Accepted: 12 February 2019 / Published online: 20 February 2019


**Abstract**
The easy implementation of replay attacks by a fraudster poses a severe threat to automatic speaker verification (ASV) technology than the other spoofing attacks like speech synthesis and voice conversion. Replay attacks refer to an attack by a fraudster to get illegitimate access to an ASV system by playing back the speech sample collected from genuine target speaker. The significant cues that can differentiate between genuine and replay recordings are channel characteristics. To capture these characteristics, one need to extract features from the spectrum, which should have high spectral and temporal resolutions. Zero time windowing (ZTW) analysis of speech is one such time-frequency analysis technique, which results in high spectral and temporal resolution spectrum at each sampling instant. In this study, new features are proposed by applying cepstral analysis to ZTW spectrum. Experiments are performed on two publicly available replay attack databases namely BTAS 2016 and ASVspoof 2017. The first set of experiments are conducted using Gaussian mixture models to evaluate the potential of proposed features. Performance of the proposed system in terms of half total error rate is 0.75% and in terms of equal error rate is 14.75% on BTAS 2016 and ASVspoof 2017 evaluation sets respectively. A score level fusion is performed by using proposed features with previously proposed single frequency filtering cepstral coefficients. This fused result outperformed the previously reported best results on these two datasets.

**Keywords** Automatic speaker recognition · Spoofing counter measures · Replay attacks · Single frequency filtering · Zero time windowing · Gaussian mixture models · Deep neural networks

## 1 Introduction

Automatic speaker verification (ASV) is the task of accepting or rejecting the identity claim of the speaker (Furui 1981). The speech signal is processed to extract speaker-specific features for building the speaker models and later use the same for testing (Pati and Prasanna 2013). The speaker-specific characteristics that are embedded in the speech signal are represented by short-term spectral, prosodic and high-level idiolectal features (Wu et al. 2015). Short-term spectral features are extracted from short segments of speech (20–30 ms), which are the acoustic correlates of voice timbre. Prosodic features are extracted from longer segments, which typically represents the speaking style of the speaker. High-level features are extracted from a lexicon to represent speaker behaviour. In case of replay attacks as shown in Fig. 1, a fraudster tries to get illegitimate access to the ASV system by playing back the pre-recorded genuine speaker voice (Wu et al. 2015). As the fraudster is using the recorded genuine speaker voice, the short-term spectral, prosodic and idiolectal features are similar in both cases. The noticeable change in genuine and replay speech can be observed in channel characteristics. However, as advanced ASV technology uses channel compensation techniques, it is easy to fool an ASV system with replay speech. As there is no other speech technology required for implementing a system for replay attacks and also because of widely available high-quality low cost recording devices, it poses a severe threat to ASV technology.

Initial studies of replay attacks on ASV technology are reported in (Villalba and Lleida 2011a; Wang et al. 2011; Villalba and Lleida 2011b; Shang and Stevenson 2010) these works are carried out on their own onsite data-sets. Recently

✉ K. N. R. K. Raju Alluri
raju.alluri@research.iiit.ac.in

Anil Kumar Vuppala
anil.vuppala@iiit.ac.in

[1] Speech Processing Laboratory, KCIS, International Institute of Information Technology, Hyderabad, India

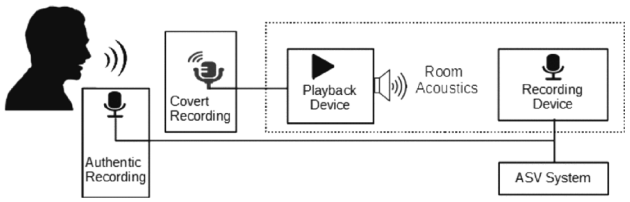


**Fig. 1** An example of replay spoofing. Reproduced with permission from Kinnunen et al. (2017b)

the ASV community has come with a wide range of replay attack databases namely, AVspoof[1] Korshunov et al. (2016) and ASVspoof 2017[2] Kinnunen et al. (2017a), with a common protocol to compare the methodologies across researchers to provide a generic solution to develop countermeasures for replay attacks. Several studies conducted on AVspoof data-set are reported in (Korshunov et al. 2016; Korshunov and Marcel 2016; Todisco et al. 2017; Paul et al. 2017; Alluri et al. 2017a) and on ASVspoof 2017 are reported in (Font et al. 2017; Patil et al. 2017; Alluri et al. 2017b; Cai et al. 2017; Chen et al. 2017; Jelil et al. 2017; Ji et al. 2017; Lavrentyeva et al. 2017; Nagarsheth et al. 2017; Li et al. 2017; Wang et al. 2017; Witkowski et al. 2017). The countermeasures for replay attacks broadly fall into four categories (Sahidullah et al. 2018). First is based on challenge-response approach (Kinnunen et al. 2016). The second one is based on template matching (Shang and Stevenson 2010). The third one is by modeling the acoustic characteriation of genuine and replay speech (Korshunov et al. 2016; Todisco et al. 2017). The final one is based on voice liveness detection (Sahidullah et al. 2018; Shiota et al. 2016). All the works reported on AVspoof and ASVspoof 2017 will fall into the third category.

Recent countermeasures developed against replay attacks have resulted in a excellent performance gain by employing unconventional features (Korshunov et al. 2016; Todisco et al. 2016, 2017), which are not generally used in ASV technology. To capture the channel characteristics in the speech signal, one has to give more importance to low signal-to-noise-ratio (SNR) instants than high SNR instants. In the case of conventional block processing based techniques, the resulting spectrum is the average characteristics of high SNR and low SNR instants because of this mechanism one cannot capture the channel characteristics from these spectra effectively. Hence, there is a need for efficient spectrum estimation techniques. With this motivation, recently we have extracted the features from single frequency filtering (SFF) (Aneeja and Yegnanarayana 2015) based spectrum, which captures the instantaneous spectral changes with high spectral and temporal resolution and the results were reported in Alluri et al. (2017a, b). In the similar lines of thought, for successful countermeasures, designing new feature representations has produced better results than with complex classifiers (Sahidullah et al. 2015). As an extension to our previous work (Alluri et al. 2017a, b), in this study, we use the features extracted from the zero time windowing (ZTW) (Bayya and Gowda 2013) spectrum as countermeasures for replay attack detection. In case of zero time windowing spectrum, the spectral representation is computed at each instant rather than for each processing block as in case of conventional spectrum estimation techniques. Cepstral analysis is performed on the ZTW spectrum to get zero time windowing cepstral coefficients (ZTWCC). Both SFF and ZTW based time-frequency analysis methods provide high spectro-temporal resolution. More specifically SFF results in high spectral resolution with a moderate temporal resolution whereas in ZTW the temporal resolution is high with reasonable spectral resolution. The method of extracting SFF spectrum is complementary to that of ZTW based spectrum, i.e., in case of SFF, the spectrum is obtained by filtering the signal at each frequency, whereas in case of ZTW, the spectrum is obtained by computing the group delay of highly decaying windowed signal at each sample. In the literature, it can be observed that the fusion of different subsystems with complementary characteristics gave significant improvement in the results. In this study, the fusion of SFF and ZTW based subsystems are explored. Here, we hypothesize that the features extracted from these two spectra catch complementary artifacts of channel pattern. From the experimental results which are evaluated on two standard data-sets, the score level fusion of these two subsystems have outperformed the individual systems, and the results are even comparable to other successful countermeasures. Further experiments are carried out to explore different deep neural network architectures for replay attack detection.

The remaining paper is organized as follows. Section 2 describes the prior works on countermeasures for replay attack detection. The motivation for high spectro-temporal resolution features is presented in Sect. 3. In Sect. 4, extraction of ZTWCC features is presented. The experimental setup used in this study is described in Sect. 5. In Sect. 6, results and discussion are presented. Finally, in Sect. 7, we discuss the summary and future scope of the studies.

## 2 Previous works on countermeasures for replay attack detection

This section presents a brief review of previous works related to replay spoofing countermeasures for automatic speaker verification. In the literature, countermeasures

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

[2] http://www.spoofingchallenge.org.

**Table 1** Description of BTAS 2016 Corpus

| Data type | | # Train | # Dev | # Test |
|---|---|---|---|---|
| Genuine | | 4973 | 4995 | 5576 |
| RE-LP-LP | R1 | 700 | 700 | 800 |
| RE-LP-HQ-LP | R2 | 700 | 700 | 800 |
| RE-PH1-LP | R3 | 700 | 700 | 800 |
| RE-PH2-LP | R4 | 700 | 700 | 800 |
| SS-LP-LP | R5 | 490 | 490 | 560 |
| SS-LP-HQ-LP | R6 | 490 | 490 | 560 |
| VC-LP-LP | R7 | 17400 | 17400 | 19500 |
| VC-LP-HQ-LP | R8 | 17400 | 17400 | 19500 |
| RE-PH2-PH3 | R9 | – | – | 800 |
| RE-LPPH2-PH3 | R10 | – | – | 800 |
| All attacks | | 38580 | 38580 | 44920 |

*RE* replay, *LP* for laptop, *HQ* means high quality speakers used during replay, *PH1* is samsung Galaxy S4 phone, *PH2* is iPhone 3GS, *PH3* is iPhone 6S, *VC* means voice conversion and SS means speech synthesis

against replay attacks based on the acoustic characterization of channel information majorly fall into the following three categories.

## 2.1 Works on on-site datasets

The first study on channel pattern noise based countermeasures is proposed in Wang et al. (2011). In this study, a support vector machine (SVM) is used to model the channel pattern features extracted from the de-noising filter and statistical frames. Experiments are conducted on authentic and playback speech database (APSD). These countermeasures have helped the ASV system to reduce the equal error rate (EER) by 30%.

## 2.2 Works on BTAS 2016 corpus

### 2.2.1 Database

BTAS 2016 corpus (Korshunov et al. 2016) is a subset of AVspoof (Ergünay et al. 2015). This corpus contains three non-overlapping subsets: train, development, and test. Each subset is further divided into two main parts: (i) genuine data, (ii) different replay attacks. Training and development data contain a similar type of attacks (R1–R8) which are known as known attacks. Whereas in the test data there are two additional attacks (R9 and R10) present in addition to the attacks present in train and development data-sets, these attacks are known as unknown attacks. The number of utterances in each type of attack is given in Table 1. Detailed information of the database can be found in Korshunov et al. (2016).

### 2.2.2 Countermeasures

BTAS 2016 speaker anti-spoofing challenge (Korshunov et al. 2016), introduced the first standard data-set for replay attacks with a common evaluation metric, i.e., half total error rate (HTER). Several studies have been conducted on BTAS 2016 corpus. A brief description of the top four performing systems is detailed below.

- SJTUSpeech (Korshunov et al. 2016): based on traditional 39-dimensional cepstral mean-variance normalisation (CMVN) along with normalized perceptual linear predictive (PLP) features, a seven-layer DNN and four-layer bidirectional long short-term (BLSTM) are used as classifiers.
- IITKGP-ABSP (Korshunov et al. 2016): based on score level fusion of two Gaussian mixture model (GMM) based sub-systems. One subsystem is built using mel frequency cepstral coefficients (MFCCs) and other using inverted MFCC (IMFCC).
- CQCC-SDA (Todisco et al. 2017): constant Q cepstral coefficients (CQCC) are coupled with GMM.
- SFFCC-SDA (Alluri et al. 2017a): based on single frequency filtering cepstral coefficients (SFFCC) coupled with GMM.

These results are reported in Table 2. First two systems in Table 2 are the results obtained during the challenge period and the last two systems are evaluated after challenge. The IITKGP-ABSP team submission is the best performing system during the challenge period. Our previous work (Alluri et al. 2017a), based on SFFCC features coupled with GMM is the state-of-the-art work on BTAS 2016 corpus.

**Table 2** Individual attack results (in % HTER) of different systems on BTAS test data set

| System | Known attacks | Unknown attacks | All attacks |
|---|---|---|---|
| SJTUSpeech (Korshunov et al. 2016) | 2.08 | 10.46 | 2.20 |
| IITKGP-ABSP (Korshunov et al. 2016) | 0.98 | 14.75 | 1.26 |
| CQCC-SDA (Todisco et al. 2017) | 0.29 | 0.29 | 0.67 |
| SFFCC-SDA (Alluri et al. 2017a) | 0.04 | 0.04 | 0.05 |

**Table 3** ASVspoof 2017 data description

| Subset | # Spk | # Replay sessions | # Replay config | # Utterances | |
|---|---|---|---|---|---|
| | | | | Genuine | Replay |
| Train | 10 | 6 | 3 | 1507 | 1507 |
| Dev | 8 | 10 | 10 | 760 | 950 |
| Test | 24 | 161 | 57 | 1298 | 12008 |
| Total | 42 | 177 | 61 | 3565 | 14465 |

## 2.3 Works on ASVspoof 2017 corpus

### 2.3.1 Database

ASVspoof 2017 database is a replayed version of RedDots. This database contains three mutually exclusive subsets: train, development, and evaluation. The development data should be used for system parameters tuning, and the results should report on evaluation data. Here replay configuration comprises of environment (E), playback device (P) and recording device (R), which are provided in the meta-data provided by the ASVspoof 2017 challenge organizers. The number of utterances in each type of attack is given in Table 3. Further details on this database can be found in Kinnunen et al. (2017a) and Delgado et al. (2018). Similar to the BTAS 2016 database, the ASVspoof 2017 evaluation data will contain unknown replay configurations to test the developed system generalization capability.

### 2.3.2 Countermeasures

More diversified publicly available database on replay attacks was introduced in ASVspoof 2017 challenge (Kinnunen et al. 2017a), with a common protocol and EER as an evaluation metric. Several researchers have submitted their countermeasures (Kinnunen et al. 2017a). CQCC coupled with GMM was provided as the baseline system. A brief description of top performing systems is detailed below.

- STC-2017 (Lavrentyeva et al. 2017): score level fusion of four sub-systems with a normalized Log power magnitude spectrum from constant Q transform and Fast Fourier Transform as features and different classifiers like convolution neural networks, recurrent neural networks and support vector machines are used.
- Biometric Vox (Font et al. 2017): based on score level fusion of two Gaussian mixture model (GMM) based sub-systems. One subsystem is built using rectangular frequency cepstral coefficients (RFCC) and other using linear frequency cepstral coefficients (LFCC).

**Table 4** Performance (in % of EER) comparision for different systems on ASVspoof 2017 database

| | Dev | Eval |
|---|---|---|
| STC-2017 (Lavrentyeva et al. 2017) | 3.93 | 6.73 |
| Biometric Vox (Font et al. 2017) | – | 10.52 |
| IIT-G (Jelil et al. 2017) | 5.31 | 13.95 |

- IIT-G (Jelil et al. 2017): based on score level fusion of several subsystems based on different features like source, instantaneous frequency and cepstral features coupled with the GMM classifier.

Evaluation results are reported in Table 4. Along with these works several other works have been explored in ASVspoof 2017 challenge. A brief description of these works are described here. Several researchers used the variant of CQCC as a feature and explored different classifiers such as DNN (Nagarsheth et al. 2017), SVM (Wang et al. 2017), convolution neural networks (CNN) (Lavrentyeva et al. 2017), Residual networks (Chen et al. 2017), i-Vector (Ji et al. 2017), fully-connected DNN (FDNN) (Cai et al. 2017). In Font et al. (2017), several features are explored using GMM as classifiers. In Alluri et al. (2017b), SFFCC coupled with GMM is studied. In Li et al. (2017), speaker identity, speech content and playback and recording device variabilites are studied using F-ratio probing tool. High-frequency features are explored in Witkowski et al. (2017). Fusion of source, instantaneous frequency, and cepstral features are explored in Jelil et al. (2017). In Patil et al. (2017), new features named variable length Teager energy operator energy separation algorithm instantaneous features (VESA-IFCC) are proposed.

# 3 Motivation for high spectro-temporal resolution spectrums

This section describes the motivation behind the use of a sample-based spectrum for replay attack detection. The spectrograms of genuine and replay recordings of the same utterance are presented in Fig. 2. It can be observed from the Fig. 2 that is the significant differences in energy distribution can be observed in non-speech regions and also at all frequencies. To capture these channel characteristics present in the speech signal, one has to give more importance to the low signal to noise ratio (SNR) instants than the high SNR instants. The majority of conventional feature extraction techniques employed in ASV technology uses block processing to represent time-frequency distribution

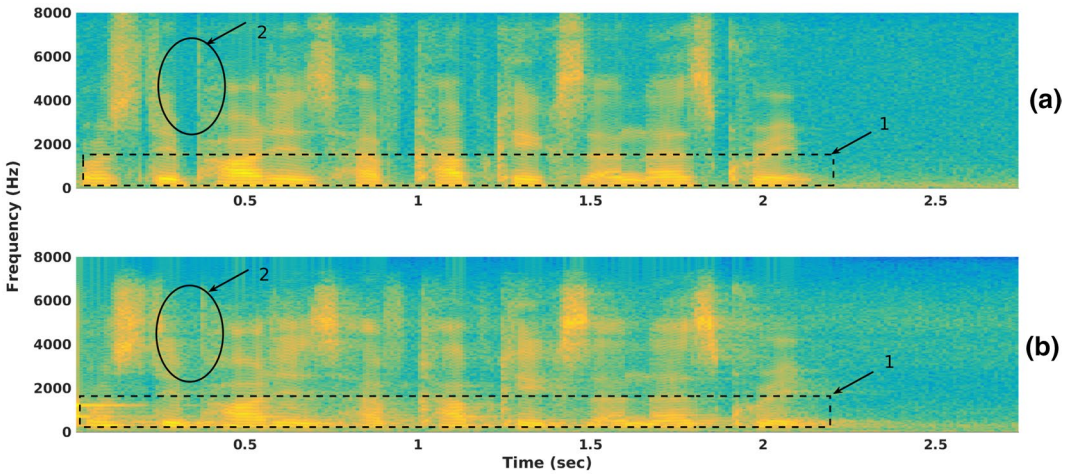


**Fig. 2** Spectrograms of **a** genuine and **b** replay utterances of same speaker and same phrase "Birthday parties have cup cakes and ice creams"

present in the signal. Because of the block processing, the resulting spectrum is the average characteristics of high SNR and low SNR instants because of this mechanism one cannot capture the channel characteristics from these kinds of techniques effectively. To overcome this issue, recently a new countermeasure is proposed for spoofing, which is based on constant Q transform, which provides high spectral resolution at low frequencies and high temporal resolution at high frequencies. The CQCC based countermeasure can differentiate moderate recordings but lacks its performance in high-quality replay recordings. To capture the cues that are more specific to replay attacks, we need a sample based spectrum with high resolution then, we can selectively pick low SNR instants from each small segment of speech. The sample-based spectrum can be computed either by filtering the frequency shifted signal at each frequency or by my multiplying a highly decaying window at each sample. By filtering method, the resultant spectrum will have the high spectral resolution at all frequencies, similarly, in other case high temporal resolution is obtained at every time instant. Single frequency filtering and zero time windowing techniques are such techniques, which results in high spectral and temporal resolutions respectively. More specifically SFF results in high spectral resolution with moderate temporal resolution where as in ZTW the temporal resolution is high with moderate spectral resolution.

Spectrograms computed using conventional short-time Fourier transform (STFT), constant Q transform (CQT), single frequency filtering (SFF) and zero time windowing (ZTW) of the same utterance are presented in Fig. 3. As mentioned earlier the lack of resolution at different frequencies can be observed in Fig. 3a similarly high spectral resolution at low frequencies and high temporal resolution at high frequencies can be observed in Fig. 3b. Whereas a high spectral resolution can be observed at each frequency in case of the spectrograms obtained from SFF and ZTW techniques can be observed in Fig. 3c, d respectively.

In our previous works, we have proposed SFFCC (Alluri et al. 2017a, b) which performed competitively with many other features. In this work, we want to propose a new feature set based on ZTW analysis, and as ZTW analysis is complementary to that of SFF analysis, we want to investigate this by fusing the results of both systems. The main idea is to come up with more specific handcrafted features which can distinguish replay recordings from genuine recordings with a simple classifier like GMM.

## 4 Zero time windowing cepstral coefficients

Motivated with the study of features extracted from high temporal and spectral resolution spectrums are useful for replay attack detection (Alluri et al. 2017a). In this study, the features are extracted from zero time windowing/liftering (ZTW/ZTL) (Bayya and Gowda 2013) based method of speech analysis. The primary goal of ZTW is to extract spectral information with high spectral and temporal resolution at any instant of time. The conventional cepstral analysis is performed on the ZTW spectrum to get zero time windowing cepstral coefficients (ZTWCC). The sequence of steps to extract ZTWCC are presented in Fig. 4 and are as follows.

1. Consider differenced signal $s[n]$ at the sampling frequency of $f_s$ Hz. In this study $f_s$ =16 kHz.

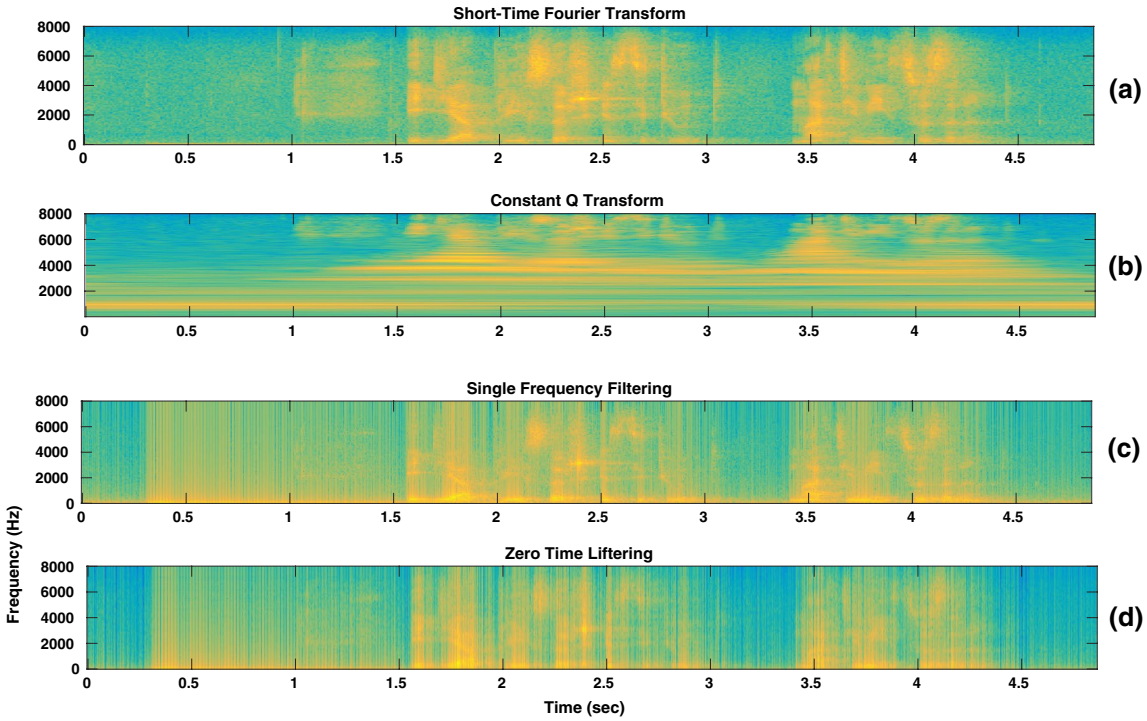


**Fig. 3** Spectrograms of the utterance "IDIAP research institute martigny switzerland" for a male speaker in AVspoof database. In **a** spectrogram is computed with short-time fourier transform where as in (**b–d**) it is computed with constant Q transform, single frequency filtering and zero time liftering techniques respectively

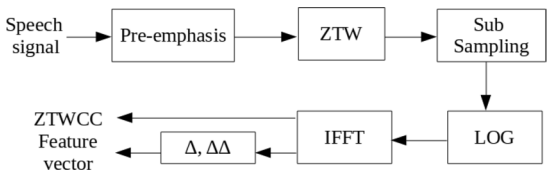


**Fig. 4** Block diagram of zero time windowing cepstral coefficients extraction

2. Consider $M$ samples of the signal, starting from an arbitrary reference set at $n = 0$. That is $s[n]$ is defined for $n = 0, 1, ..., M - 1$.
3. Compute the windowed signal $x[n] = s[n]w[n]$, for $n = 0, 1, ..., M - 1$, where $w[n]$ is defined by

$$w[n] = \begin{cases} 0 & n = 0 \\ 1/(4\sin^2(\pi n/M)) & n = 1, 2, ..., M-1 \end{cases} \quad (1)$$

ZTL spectrum computation involves the windowing speech signal using $w[n]$ with one sample shift.

4. Append $x[n]$ with $(N - M)$ zeros and compute numerator of group delay (NGD). Here N should be a very large value than M, so that spectral features are visible in NGD function because of increased number of samples in frequency domain. The NGD function is given by

$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], \quad k = 0, 1, ..., N-1 \quad (2)$$

where $X[k] = X_R[k] + jX_I[k]$ is the $N$-point DFT of the sequence $x[n]$, and $Y[k] = Y_R[k] + jY_I[k]$ is the $N$-point DFT of the sequence $y[n] = nx[n]$. Note that in these notations $X[k] = X(\omega)|_{\omega=2\pi k/N}$, $Y[k] = Y(\omega)|_{\omega=2\pi k/N}$, and $g[k]$ is obtained through $X[k]$ and $Y[k]$.

5. In order to get the better spectral features sign reversed double differenced NGD spectrum is computed.

$$dg(k) = -(diff(diff(g(k)))) \quad (3)$$

6. In order to further highlight spectral features Hilbert envelope (HE) of $dg(k)$ is computed.

$$v(k) = |dg(k)| \quad (4)$$

7. Repeat step 2–6 for the entire signal with one sample shift. So we get v(k,p),where p = 0 to P − 1 , P is total number of samples in signal.
8. Apply cepstral analysis to get $c(k, p)$

$$c[k, p] = IFFT(log(v[k, p])) \quad (5)$$

These cepstral features are termed as zero time windowing cepstral coefficients (ZTWCC).

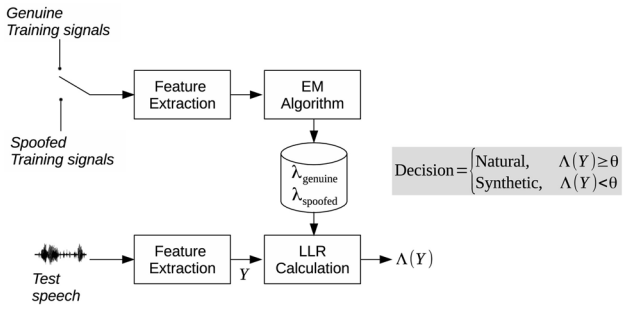


**Fig. 5** GMM-based spoofing detection system. Reproduced with permission from Hanili (2018)

## 5 Experimental setup

### 5.1 Features

Experiments are performed using proposed ZTWCC along with recently proposed SFFCC features. Parameters considered for extracting these features are mentioned below.

#### 5.1.1 SFFCC parameters

The parameters considered for SFFCC computation are taken from Alluri et al. (2017a, b), i.e., The amplitude envelope is computed for each 15.6 Hz spacing, which results 513 frequencies in the range of 0–$f_s/2$ , here $f_s$ = 16,000 Hz. With the intuition of low SNR instants are useful for replay attack detection, the SFFCC features are extracted at each low SNR instant within 10 ms segment. Thirty-dimensional features are considered for experimentation.

#### 5.1.2 ZTWCC parameters

For ZTWCC computation during step 2, M is considered as 80 samples, which refers to 5 ms of the speech segment. N is considered as 512 to get high resolution. Once we get the sample based spectrum as shown in Fig. 3. We applied sub-sampling similar to that of SFFCC extraction (Alluri et al. 2017a) i.e., the Cepstral features are extracted from each 10 ms segment by selecting the low SNR instants. Several experiments are conducted to choose the dimensionality of ZTWCC features, and the results are reported in Table 5.

### 5.2 Classifier

In this study, two subsystems are considered. The First subsystem is based on SFFCC and the second one is based on ZTWCC. For these two subsystems, GMM is used as a classifier. For each subsystem, genuine ($\lambda_{genuine}$) and spoof ($\lambda_{spoofed}$)

**Table 5** Performance (in % of EER) for ZTWCC with different configurations on ASVspoof 2017 development dataset using GMM classifier

| | Dimensions | | | |
|---|---|---|---|---|
| | 13 | 20 | 30 | 40 |
| S | 12.41 | 11.19 | **9.64** | 10.32 |
| D | 10.83 | 10.59 | 12.06 | 13.42 |
| A | 15.94 | 15.98 | 18.98 | 15.96 |
| SD | 7.21 | 6.55 | 5.98 | 6.86 |
| SA | 8.34 | 9.58 | 10.31 | 9.86 |
| DA | 9.90 | 10.22 | 10.65 | 10.32 |
| SDA | 6.86 | 6.33 | **5.20** | 6.55 |

Bold values represents the best performing numbers across the sysems

GMMs are built after the feature extraction with 512 mixture components. Expectation and maximization (EM) algorithm are used to estimate the parameters of GMM. The block diagram for GMM based spoof detection is presented in Fig. 5.

For the given test utterance Y, the log-likelihood score is computed in the following manner,

$$Score(Y) = llk(Y|\lambda_{\text{genuine}}) - llk(Y|\lambda_{\text{spoofed}}) \quad (6)$$

where Y = {$y_1$, $y_2$,…, $y_T$} here, T represents the number of sub-sampling instances. where,

$$llk(Y|\lambda) = (1/T)\sum_{t=1}^{T} logp(y_t/\lambda) \quad (7)$$

represents the average likelihood of Y given model $\lambda$.

### 5.3 Evaluation metrics

Evaluation metrics are considered according to the challenge protocol. For the experiments conducted on BTAS 2016 database, the evaluation results are reported in terms of HTER, whereas for ASVspoof 2017 database the results are reported in terms of EER. BOSARIS toolkit (2013) is used to calculate both EER and HTER.

## 6 Results and discussion

The experiments are carried out in three subsets. The first set of experiments are conducted on ASVspoof 2017 development data to fix the feature combination and dimensionality of ZTWCC feature. The second set of experiments are carried out on ASVspoof 2017 and third set on BTAS 2016 corpus.

**Table 6** Performance (in % of EER) for ZTWCC and SFFCC based systems using CMVN on ASVspoof 2017 database

| | Without CMVN | With CMVN |
|---|---|---|
| SFFCC-SDA | | |
| Dev | 4.44 (Alluri et al. 2017b) | 4.39 |
| Eval | 31.13 (Alluri et al. 2017b) | 11.60 |
| ZTWCC-SDA | | |
| Dev | 9.56 | 5.20 |
| Eval | 29.72 | 14.50 |

**Table 7** Performance (in % of EER) comparision for different systems on ASVspoof 2017 database

| | Dev | Eval |
|---|---|---|
| STC-2017 (Lavrentyeva et al. 2017) | 3.93 | 6.73 |
| SFFCC-SDA | 4.39 | 11.60 |
| ZTWCC-SDA | 5.20 | 14.50 |
| Fused system | **3.10** | **6.24** |

Bold values represents the best performing numbers across the sysems

### 6.1 Different combinations of ZTWCC

To choose the right combination and feature dimension of ZTWCC for replay attacks, several experiments are conducted on ASVspoof 2017 development data-set, and the results are reported in Table 2. Here, the first column refers to several combinations of static(S), dynamic (D) and acceleration (A) coefficients. The second row refers to the number of static coefficients.

Results from the Table 5, suggests us that the higher (30) dimensional features are useful for replay attack detection and the static appended with dynamic coefficients are more useful than the other combinations. Further experiments are carried out using 30 dimensions.

### 6.2 Results on ASVspoof 2017 dataset

Our previous work submitted to ASVspoof 2017 (Alluri et al. 2017b), based on SFFCC does not employ any cepstral mean-variance normalization (CMVN) techniques. In this study, CMVN has applied for SFFCC based features as well as for the proposed ZTWCC features, and the results are reported in Table 6.

From the results, it can be observed that a simple CMVN applied on these features have resulted in a considerable performance improvement. On ASVspoof 2017 evaluation data ZTWCC with CMVN results in an EER of 14.50%, whereas without CMVN it is 29.72%. As evaluation data-set contains several unknown configurations, there will be a bias introduced in these recordings. As we are not modeling these unknown configurations, without CMVN we are unable to get better results. With simple CMVN this bias can be compensated, and the same is reflected in accordance with the results reported in Delgado et al. (2018). Further experimental results on ASVspoof development and evaluation data-sets are reported in Table 7. The results of top performed system in ASVspoof 2017 were taken from Lavrentyeva et al. (2017).

From the results in Table 7, it can be observed that the SFFCC and ZTWCC based systems have resulted in a competitive performance with each other, whereas they are behind the top performing system. The top-performing system is the fusion of several subsystems with complex classifiers. Score-level fusion of these two subsystems are explored, and the results are reported in the last row of Table 7. Interestingly, the fused system result is performing better than the top performing system. This result ensures that these features contain complementary information which is useful for the detection of replay attacks. To better understand the complementary nature of SFFCC and ZTWCC, the results on the evaluation set are presented in terms of quality of replay configurations and the results are reported in Table 8.

As described in Sect. 2.3.1, Each replay configuration (RC) comprises of three components namely, environment (E), playback device (P) and recording device (R). These replay configurations are ranked in terms of the threat they present to automatic speaker verification. A poor quality RC should pose the least threat to ASV and also it should be easy to detect with the spoofing countermeasures. Similarly, a high-quality RC should be hard to detect and also it poses a severe threat to ASV technology. To correlate these reasons, the results are analyzed in terms of the quality of RC.

From the results in Table 8, it can be seen that for each countermeasure the EER increases from low quality to high quality in each component of replay configuration. From

**Table 8** Performance (in % of EER) comparison in-terms of difficulty level in replay configuration (environment, playback device and recording device) for different systems on ASVspoof 2017 evaluation dataset

| System | Environment | | | Play back device | | | Recording device | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Low | Medium | High | Low | Medium | High | Low | Medium | High | |
| SFFCC-SDA | 9.55 | 10.37 | 16.50 | 8.29 | 9.35 | 13.83 | 10.66 | 14.23 | 11.47 | 11.60 |
| ZTWCC-SDA | 9.65 | 13.40 | 23.25 | 10.93 | 13.49 | 17.11 | 14.11 | 13.93 | 15.09 | 14.50 |
| Fused system | 5.12 | 6.00 | 10.16 | 4.91 | 5.18 | 7.95 | 6.26 | 7.64 | 6.96 | 6.24 |

**Table 9** Individual attack results (in % HTER) of different systems on BTAS test data set

| System | Known attacks | | | | | | | | Unknown attacks | | All attacks |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Pool |
| CQCC-SDA (Todisco et al. 2017) | 0.29 | 0.29 | 0.29 | 0.29 | 0.29 | 0.29 | 0.29 | 0.29 | 0.29 | 23.92 | 0.67 |
| SFFCC-SDA (Alluri et al. 2017a) | 0.16 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.29 | 0.05 |
| ZTWCC-S | 0.51 | 0.26 | 0 | 1.19 | 0 | 0 | 0 | 0 | 0.59 | 3.53 | 0.30 |
| ZTWCC-DA | 34.50 | 13.29 | 48.28 | 49.67 | 4.18 | 1.05 | 0.39 | 0.53 | 49.98 | 25.41 | 7.29 |
| ZTWCC-SDA | 1.10 | 1.63 | 0.40 | 4.40 | 0 | 0 | 0 | 0 | 0.62 | 2.35 | 0.75 |
| Fused system | **0.05** | 0.22 | 0 | **0.03** | 0 | 0 | 0 | 0 | **0** | **0.50** | **0.02** |

Bold values represents the best performing numbers across the sysems

this result, we can say that detecting high-quality recording is very hard to that of low-quality recordings. In majority cases, ZTWCC is underperforming SFFCC, especially in high-quality cases, whereas in case of the fused system the proposed system is performing far better than their individual counterparts. One more interesting observation from this result is out of three components environment plays a crucial role in RC. A high-quality environment recording is tough to detect, and it can be supported with the EERs of SFFCC, ZTWCC in case of a high-quality environment case. These results support that ZTWCC and SFFCC are capturing complementary artifacts in detecting replay attacks.

### 6.3 Results on BTAS 2016 dataset

According to the BTAS 2016 challenge protocol, all the results on this data-set are reported in terms of HTER. Initial experiments are performed on development data set to find the threshold that is later used to calculate HTER on the evaluation data set. The results on development set are not reported here because all the systems considered here resulted in almost zero EER on development data. The Individual attack results on the evaluation set are reported in Table 9. The corresponding attack details are adapted from Alluri et al. (2017a). Proposed ZTWCC based system results are compared with previously reported successful countermeasures (Todisco et al. 2017; Alluri et al. 2017a). CQCC-SDA system is able to detect the R1–R9 attacks but it is poorly detecting the R10. From the experimental results reported in Table 9, it can be observed that ZTWCC-SDA and ZTWCC-S based systems resulted in a competitive performance with CQCC-SDA (Todisco et al. 2017) based system and a little lower performance to that of SFFCC-SDA (Alluri et al. 2017a) based system. Similar score-level fusion of ZTWCC-SDA and SFFCC-SDA based sub-systems is employed on BTAS 2016 database, and the result is reported in the last row of Table 9. Similar to the result of ASVspoof 2017, the fused system outperformed the individual subsystem results and became the new state-of-the result on this data-set. From the individual attack results, it can be observed that ZTWCC-DA system can detect replayed voice conversion and speech synthesis based attacks (R5-R8). This is because of VC, and SS based speech signals lack the long-term temporal dynamics. The reason to choose ZTWCC-SDA based system for the fusion with SFFCC based system is that of generalization capability of ZTWCC-SDA than the ZTWCC-S, i.e., ZTWCC-SDA is better performing in case of unknown scenarios than the ZTWCC-S.

Score-level fusion of ZTWCC and SFFCC based subsystems have resulted in great performance improvement, and the fused system result became the new state-of-the-art result on both ASVspoof 2017 and BTAS 2016 corpora. BTAS 2016 corpus comprises very less number of RC and most of them are of low quality, because of this reason the results are very impressive than that of ASVspoof 2017 dataset.

### 6.4 Results using deep neural networks

The motivation of this study is to explore a suitable feature for replay attack detection. The main reason for selecting GMM for most of the studies is its ability to capture feature

**Table 10** Performance (in % of EER) comparision for different systems on ASVspoof 2017 database

| | SFFCC | | ZTWCC | |
|---|---|---|---|---|
| | Dev | Eval | Dev | Eval |
| GMM | 4.39 | 11.60 | 5.20 | **14.50** |
| DNN-1L | 4.08 | 13.85 | 10.21 | 16.23 |
| DNN-2L | 3.92 | 13.16 | 9.27 | 16.08 |
| DNN-3L | 4.01 | 11.36 | 10.25 | 16.80 |
| DNN-4L | 4.57 | **11.00** | 12.34 | 18.37 |
| DNN-6L | 4.68 | 12.80 | 10.88 | 16.14 |

Bold values represents the best performing numbers across the sysems

**Table 11** Performance (in % of EER) for DNN based system on ASVspoof 2017 database

| | Dev | Eval |
|---|---|---|
| SFFCC-DNN | 4.57 | 11.00 |
| ZTWCC-DNN | 9.27 | 16.08 |
| Fused system | **3.70** | **8.8** |

Bold values represents the best performing numbers across the sysems

distribution with low dimensionality features. In this section, we have explored the deep neural networks on SFFCC and ZTWCC, and the results are reported in Table 10. The first column in Table 10 represents the best architecture with the mentioned number of layers, i.e., DNN-2 L means the best-performed system on development data of ASVspoof 2017. Here, the best performing system is chosen by conducting several experiments with different hyper-parameters like a number of nodes in each layers learning rate, learning algorithm and so on.

From the results in Table 10, it can be observed that the best performing DNN based system is slightly better than the best performing GMM based system in case of SFFCC. Whereas in the case of ZTWCC is GMM based system is performing better than the DNN based system. The feature distribution in case of ZTWCC is better captured in the case of GMM rather than that of DNN. Here our intention is to say the explored features are able to differentiate genuine and replay attacks effectively with DNN architecture also and the complementary nature is observed in DNN case also. The reason for the depreciation of result in case of DNNs may the feature selected here is of low dimension, and one more reason is as the evaluation set is almost different from development and training sets, so the network is more biased to the development set.

The score-level fusion of SFFCC and ZTWCC based best DNN system is performed, and the result is reported in Table 11. Also, we can observe the complementary nature of the features because of which the fused system is performing better than that of individual systems.

## 7 Summary and conclusions

Replay attacks pose a serious threat to speaker recognition technology due to its easily accessible nature. The major differences between genuine recordings and replay recordings lie in channel pattern characteristics. In this paper, the importance of high spectro-temporal resolution features for replay attack detection is explored. A new set of features namely, Zero time windowing based cepstral coefficients are introduced for replay attack detection. Recently proposed single frequency filter cepstral coefficients are also used in this study. Experimental analysis is carried out on ASVspoof 2017 and BTAS 2016 databases. The complementary nature of SFFCC and ZTWCC helped in detecting replay attacks efficiently than that of previously reported successful countermeasures. This complementary nature is explored by using a score-level fusion of individual subsystems. Further, the complementary nature of features is explored using deep neural network architectures. Proposed systems are the new state-of-the-art for BTAS 2016 ad ASVspoof 2017 databases. They reduce the error rate by 60% and 7% in BTAS 2016 and ASVspoof 2017 respectively.

**Acknowledgements** Authors thank Mr. Sudarsana Reddy Kadiri, and Mr. Sivanand Achanta of IIIT-Hyderabad for assistance with single frequency filtering and zero time windowing techniques. The first author would like to thank the Department of Electronics and Information Technology, Ministry of Communication & IT, Govt of India for granting Ph.D. Fellowship under Visvesvaraya Ph.D. Scheme.

## References

Alluri, K. R., Achanta, S., Kadiri, S. R., Gangashetty, S. V., & Vuppala, A. K. (2017a). Detection of replay attacks using single frequency filtering cepstral coefficients. In *Proceedings of the Interspeech 2017* (pp. 2596–2600).

Alluri, K. R., Achanta, S., Kadiri, S. R., Gangashetty, S. V., & Vuppala, A. K. (2017b). Sff anti-spoofer: Iiit-h submission for automatic speaker verification spoofing and countermeasures challenge 2017. In *Proceedings of the Interspeech* (pp. 107–111).

Aneeja, G., & Yegnanarayana, B. (2015). Single frequency filtering approach for discriminating speech and nonspeech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(4), 705–717.

Bayya, Y., & Gowda, D. N. (2013). Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. *Speech Communication*, *55*(6), 782–795.

Brümmer, N., & de Villiers, E. (2013). The BOSARIS Toolkit: Theory, algorithms and code for surviving the New DCF. arXiv preprint arXiv:1304.2865.

Cai, W., Cai, D., Liu, W., Li, G., & Li, M. (2017). Countermeasures for automatic speaker verification replay spoofing attack : On data augmentation, feature representation, classification and fusion. In *Proceedings of the Interspeech 2017* (pp. 17–21).

Chen, Z., Xie, Z., Zhang, W., & Xu, X. (2017). Resnet and model fusion for automatic spoofing detection. In *Proceedings of the Interspeech 2017* (pp. 102–106).

Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K. A., & Yamagishi, J. (2018). Asvspoof 2017 version 2.0: Meta-data analysis and baseline enhancements. In *Proceedings of the Odyssey 2018 the speaker and language recognition workshop* (pp. 296–303).

Ergünay, S. K., Khoury, E., Lazaridis, A., & Marcel, S. (2015). On the vulnerability of speaker verification to realistic voice spoofing. In *Proceedings of the BTAS* (pp. 1–6).

Font, R., Espn, J. M., & Cano, M. J. (2017). Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge. In *Proceedings of the Interspeech 2017* (pp. 7–11).

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

Hanilçi, C. (2018). Data selection for i-vector based automatic speaker verification anti-spoofing. *Digital Signal Processing*, *72*, 171–180.

Jelil, S., Das, R. K., Prasanna, S. M., & Sinha, R. (2017). Spoof detection using source, instantaneous frequency and cepstral features. In *Proceedings of the Interspeech 2017* (pp. 22–26).

Ji, Z., Li, Z.-Y., Li, P., An, M., Gao, S., Wu, D., & Zhao, F. (2017). Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof 2017. In *Proceedings of the Interspeech 2017* (pp. 87–91).

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N.,Yamagishi, J., & Lee, K. A. (2017a). The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Proceedings of the 18th annual conference of the international speech communication association* (pp. 2–6).

Kinnunen, T., Sahidullah, M., Falcone, M., Costantini, L., Hautamaki, R. G., Thomsen, D. A. L., Sarkar, A. K., Tan, Z.-H., Delgado, H., & Todisco, M., et al. (2017b). RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In *IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, 2017* (pp. 5395–5399)

Kinnunen, T., Sahidullah, M., Kukanov, I., Delgado, H., Todisco, M., Sarkar, A. K., Thomsen, N. B., Hautamäki, V., Evans, N. W., & Tan, Z.-H. (2016). Utterance verification for text-dependent speaker recognition: A comparative assessment using the reddots corpus. In *Proceedings of the Interspeech* (pp. 430–434).

Korshunov, P., & Marcel, S. (2016). Cross-database evaluation of audio-based spoofing detection systems. In *Proceedings of the Interspeech* (pp. 1705–1709).

Korshunov, P., Marcel, S., Muckenhirn, H., Gonçalves, A., Mello, A. S., Violato, R. V., Simoes, F., Neto, M., de Assis Angeloni, M., Stuchi, J., et al. (2016). Overview of BTAS 2016 speaker anti-spoofing competition. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1–6).

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. In *Proceedings of the Interspeech* (pp. 82–86).

Li, L., Chen, Y., Wang, D., & Zheng, T. F. (2017). A study on replay attack and anti-spoofing for automatic speaker verification. In *Proceedings of the Interspeech 2017* (pp. 92–96).

Nagarsheth, P., Khoury, E., Patil, K., & Garland, M. (2017). Replay attack detection using DNN for channel discrimination. In *Proceedings of the Interspeech 2017* (pp. 97–101).

Pati, D., & Prasanna, S. M. (2013). A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information. *Sadhana*, *38*(4), 591–620.

Patil, H. A., Kamble, M. R., Patel, T. B., & Soni, M. H. (2017). Novel variable length teager energy separation based instantaneous frequency features for replay detection. In *Proceedings of the Interspeech 2017* (pp. 12–16).

Paul, D., Sahidullah, M., & Saha, G. (2017). Generalization ofspoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2047–2051).

Sahidullah, M., Kinnunen, T., & Hanilçi, C. (2015). Acomparison of features for synthetic speech detection. In *Proceedings of the Interspeech* (pp. 2087–2091).

Sahidullah, M., Thomsen, D. A. L., Hautamäki, R. G., Kinnunen, T., Tan, Z.-H., Parts, R., et al. (2018). Robust voice liveness detection and speaker verification using throat microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 44–56.

Shang, W., & Stevenson, M. (2010). Score normalization in playbackattack detection. In *2010 IEEE international conference on acoustics, speech and signal processing* (pp. 1678–1681).

Shiota, S., Villavicencio, F., Yamagishi, J., Ono, N., Echizen, I., & Matsui, T. (2016). Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector. In *Proceedings of the Odyssey: Speaker language recognition workshop* (Vol. 2016, pp. 259–263).

Todisco, M., Delgado, H., & Evans, N. (2016). A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In Proceedings of the Speaker Odyssey Workshop, Bilbao, Spain (Vol. 25, pp. 249–252).

Todisco, M., Delgado, H., & Evans, N. (2017). Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech and Language*, *45*, 516–535.

Villalba, J., & Lleida, E. (2011a). Detecting replay attacks from far-field recordings on speaker verification systems. In *Proceedings of the European workshop on biometrics and identity management* (pp. 274–285).

Villalba, J., & Lleida, E. (2011b). Preventing replay attacks onspeaker verification systems. In *IEEE international carnahan conference on security technology (ICCST)* (pp. 1–8).

Wang, X., Xiao, Y., & Zhu, X. (2017). Feature selection based on CQCCS for automatic speaker verification spoofing. In *Proceedings of the Interspeech 2017* (pp. 32–36).

Wang, Z.-F., Wei, G., & He, Q.-H. (2011). Channel pattern noise based playback attack detection algorithm for speaker recognition. In *International conference on machine learning and cybernetics, Guilin, 2011* (pp. 1708–1713).

Witkowski, M., Kacprzak, S, Elasko, P., Kowalczyk, K., & Gaka, J. (2017). Audio replay attack detection using high-frequency features. In *Proceedings of the Interspeech 2017* (pp. 27–31).

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, *66*, 130–153.