Novel Spectral Root Cepstral Features for Replay Spoof Detection

Prasad A. Tapkir, Ankur T. Patil, Neil Shah, and Hemant A. Patil
Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
E-mail: {prasad_tapkir, ankur_patil, neil_shah, hemant_patil}@daiict.ac.in

***Abstract*—Replay poses a greater threat to the Automatic Speaker Verification (ASV) system than any other spoofing attacks, as it neither require any specific expertise nor a sophisticated equipment. In this paper, we propose a novel countermeasure by modeling the replayed speech as a convolution of genuine speech with additional impulse responses (due to the microphone, loudspeaker, recording and replay environment). In particular, we propose the new feature set, namely, Magnitude-based Spectral Root Cepstral Coefficients (MSRCC) and Phase-based Spectral Root Cepstral Coefficients (PSRCC), that performs better than the baseline system (CQCC), on ASVspoof 2017 challenge database, which gives 29.18 % Equal Error Rate (EER) on the evaluation set. The proposed feature set detects the effect of these additional impulse responses, in the quefrency-domain. Experiments performed on evaluation set using MSRCC and PSRCC, with Gaussian Mixture Model (GMM) as a classifier gives 18.61 % and 24.35 % EER, respectively. On the other hand, Convolutional Neural Network (CNN) classifier gives 24.50 % and 26.81 % EER, respectively. The score-level fusion of MSRCC and PSRCC gives reduced EER of 10.65 % using GMM and 17.76 % using CNN classifier, indicates the complementary information captured by the proposed feature sets.**

I. Introduction

In replay attack, an attacker uses recorded speech samples of the target speaker to get an access to the Automatic Speaker Verification (ASV) system. It is very difficult to detect these attacks if high-quality recording devices are used, because they produce very similar replayed speech signal to that of the natural speech signal. In addition, it is the simplest spoofing attack as it do not require any specific knowledge in speech processing or any sophisticated computer-aided technologies [1]. In practice, we would like ASV system to be robust against variations, such as, microphone and transmission channel, intersession, acoustic noise, speaker ageing, etc. This robustness makes ASV system vulnerable to replay attacks, as it tries to *nullify* these effects and make replayed speech more close to the natural speech. Hence, we would like the system to be secure against such spoofing attacks.

One of the initial study in replay spoofed speech detection (SSD) for the text-dependent system was reported in [2]. To verify the input speech, a choice is made based on a set of N similarity scores is used. For detection of concatenated segments of speech, the study in [3] uses F_0 and Mel Frequency Cepstral Coefficients (MFCC) feature set. For text-dependent ASV system, the spectral bitmaps are used to determine whether the input speech is natural or replayed [4]. In [5], similar technique of average spectral bitmaps is used for text-independent ASV system. ASV spoof 2017 challenge is organized to develop the countermeasures of the replay attack detection on highly heterogeneous recording and replayed conditions. Baseline system uses Gaussian Mixture Model (GMM) as back-end classifier with Constant Q Cepstral Coefficients (CQCC) feature set that is based on the perceptually-motivated time-frequency transform [6]. Study reported in [7] investigated different spectral features, such as, CQCC, MFCC, etc. and their feature-level fusion. The cross-database experiments with the BATS 2016 ASVspoof development set are also reported. A new feature extraction approach, namely, Variable length Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (VESA-IFCC) is proposed to exploit the usefulness of instantaneous frequency (IF) in subband energy via Energy Separation Algorithm (ESA) [8]. The study in [9], have explored different feature set as glottal closure instants, epoch strength and the peak-to-sidelobe ratio of Hilbert envelope of the Linear Prediction (LP) residual. In [10], data augmentation along with deep Residual Network (ResNet) is used. The same study also shows experiment using Deep Neural Network (DNN), Bidirectional Long Short Term Memory (BLSTM) neural network as classifiers with CQCC features.

As replay spoofing attack affects on the high frequency spectrum, analysis using inverse-MFCC, Linear Prediction Cepstral Coefficients (LPCC), and LPCC-residual (LPCCres) features is performed in high frequency region [11]. In [12], feature selection methods are applied on mean and variance of CQCC features with Support Vector Machine (SVM) as a classifier. DNN architectures, such as Light CNN , CNN+ Recurrent Neural Network (RNN) are used along with *i*-vector+SVM, constant Q transform (CQT), and FFT as feature representation [13]. In [14], they proposed ensemble classifier set using multiple GMM, GMM mean supervector-Gradient Boosting Decision Tree (GSV-GBDT) and GSV-Random Forest (GSV-RF) classifiers. An F-ratio probing tool is used to analyze the three variability factors, i.e., speaker identity, speech content and playback and recording device [15]. In these set of experiments, it is observed that replay device contributes to *overfitting* risk. The fusion of high level features using DNN with the CQCC and High Frequency Cepstral Coefficients (HFCC) is investigated in [16]. Model fusion

using GMM, DNN, and ResNet is performed in [17]. It is found that multi-channel information in replayed speech is found in low SNR regions. The analysis is performed using Single Frequency Filtering (SFF) [18].

In the proposed method, we exploit source-filter model of the natural speech production by modeling the genuine speech signal as the convolution of excitation source (glottal airflow) and system (vocal tract) impulse response [19]. These convolutionally combined signals cannot be distinctly observed in the spectral-domain. Furthermore, the speech signal may be convolved with the other system responses, such as, transmission channel or by the flawed recording device. In the context of replay spoof modeling, we can model the spoofed speech signal as a convolution of the genuine speech with the other additional elements, such as, acoustic effects introduced by the recording device, recording environmental conditions, replayed device and acoustics of the environment where the attack takes place. We propose the new feature sets, Magnitude-based Spectral Root Cepstral Coefficients (MSRCC) and Phase-based Spectral Root Cepstral Coefficients (PSRCC), for SSD system. In these features, effect of additional elements in replayed speech is over entire quefrency-domain which is helpful in SSD task. In addition, these sets of feature, contain very high complementary information as their score-level fusion gives the significant improvement in performance as compared to the standalone MSRCC features. The proposed feature sets perform significantly better than the baseline CQCC-GMM SSD system.

II. Speech Modeling and Cepstrum

A. Speech Modeling

Let $s(n)$ be the genuine speech signal that can be modeled as a convolution of glottal airflow, $p(n)$ and vocal tract impulse response $h(n)$ [20].

$$s(n) = p(n) * h(n), \tag{1}$$

where $*$ denotes the convolution. Given genuine speech, $s(n)$, replay speech signal can be modeled as [21]:

$$r(n) = s(n) * h_{mic}(n) * a(n) * h_{spk}(n) * b(n), \tag{2}$$

where $h_{mic}(n)$ and $h_{spk}(n)$ are impulse responses of recording microphone and loudspeaker, respectively, and $a(n)$ and $b(n)$ are impulse responses of recording and replayed environments, respectively. Further, Eq. (2) can be simplified as,

$$r(n) = s(n) * N(n), \tag{3}$$

where $N(n) = h_{mic}(n) * a(n) * h_{spk}(n) * b(n)$. In this study, we aim to detect these extra convolved elements for which we can apply homomorphic signal processing techniques. Generally, there are two homomorphic techniques, namely: 1) The system that transfers convolutional vector space into additive vector space [22], and 2) The system which maps convolutional vector space into another convolutionally combined vector space [23]. The purpose of both the homomorphism is to separate the convolutionally combined signals by compressing the impulse responses w.r.t. the impulse train of the glottal pulse.

B. Logarithmic vs. Spectral Root Cepstrum

To evaluate the cepstrum, we map convolutionally combined signals, $s(n) = p(n) * h(n)$ to additively combined signals, $\hat{s}(n) = \hat{p}(n) + \hat{h}(n)$, so that we can distinctly observe the effect of impulse train $p(n)$ and impulse response of the system $h(n)$. This transformation should take place such that $\hat{p}(n)$ remains the train of pulses with similar duration as $p(n)$, but $\hat{h}(n)$ is more time-limited than $h(n)$. The cepstrum $\hat{s}(n)$ of the signal $s(n)$ can be obtained by inverse Z transform of *log* of the Z transform of the signal. Z-transform of the signal followed by *log* ensures the transformation in additive vector space. Hence, cepstrum of the replayed speech can be expressed as:

$$\hat{r}(n) = \hat{s}(n) + \hat{N}(n). \tag{4}$$

To obtain spectral root cepstrum, convolutionally combined signal, $s(n) = p(n) * h(n)$ is mapped to another convolutionally combined signal, $\check{s}(n) = \check{p}(n) * \check{h}(n)$, such that the new convolutionally combined vectors are more easily separable. In spectral root cepstrum, logarithmic operator during cepstrum computation is replaced by exponent γ. Z transform maps the convolutional vector space into multiplicative vector space to give $S(z) = P(z) \cdot H(z)$. Then, $\check{S}(z) = [P(z) \cdot H(z)]^{\gamma} = P^{\gamma}(z) \cdot H^{\gamma}(z)$,

$$\therefore \check{S}(z) = \check{P}(z) \cdot \check{H}(z). \tag{5}$$

For above set of equations, there is a one-to-one mapping between the time-domain vectors and the Z-domain vectors. In addition, there is an implicit assumption that $s(n)$ is a real and stable sequence. Taking inverse Z transform of Eq. (5), we get,

$$\check{s}(n) = \check{p}(n) * \check{h}(n), \tag{6}$$

where $\check{s}(n)$ is known to be spectral root cepstrum.

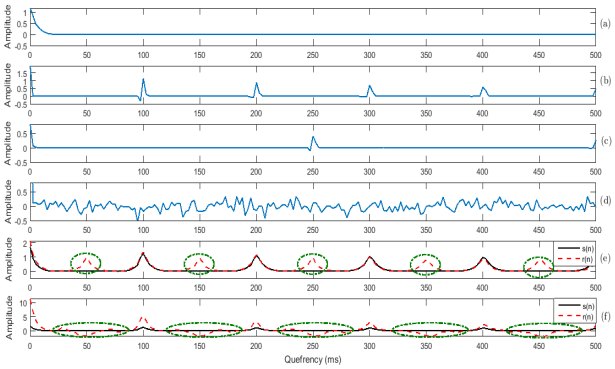

Fig. 1. Spectral root cepstrum of synthetic (a) $h(n)$, (b) $p(n)$, (c) $N(n)$ (impulsive), (c) $N(n)$ (white Gaussian noise), (e) $s(n)$ and $r(n)$ for impulsive noise, and (f) $s(n)$ and $r(n)$ for white Gaussian noise (*highlighted portion indicates effect of* $N(n)$ *component*).

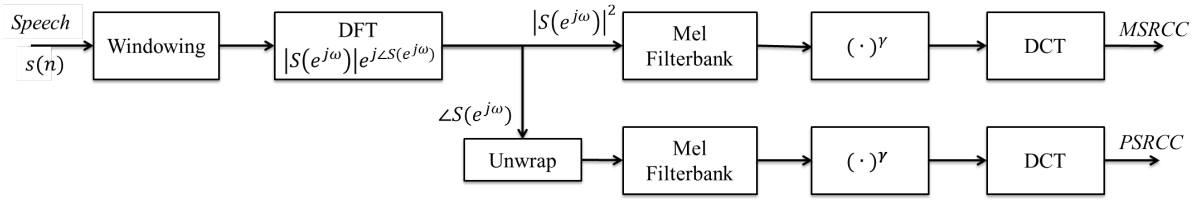

Fig. 2. Functional block diagram of proposed MSRCC and PSRCC feature extraction methodology.

Since $\hat{S}(z) = \log(S(z))$ and $\check{S}(z) = (S(z))^\gamma$, $\check{S}(z)$ is related to $\hat{S}(z)$ by,

$$\check{S}(z) = e^{\gamma \cdot \hat{S}(z)} = \sum_{n=0}^{-\infty} \frac{(\gamma \hat{S}(z))^n}{n!}, \tag{7}$$

$$\therefore \check{s}(n) = \delta(n) + \gamma \hat{s}(n) + \frac{\gamma^2}{2}\hat{s}(n) * \hat{s}(n) + \frac{\gamma^3}{3}\hat{s}(n) * \hat{s}(n) * \hat{s}(n) + ... \tag{8}$$

From Eq. (8), we can observe that even though $s(n)$ is of limited duration, $\check{s}(n)$ will be having infinite duration. For genuine speech signal $s(n)$ and spoofed speech signal $r(n)$, spectral root cepstrum can be given as,

$$\check{s}(n) = \check{p}(n) * \check{h}(n), \tag{9}$$

$$\check{r}(n) = \check{s}(n) * \check{N}(n) = \check{p}(n) * \check{h}(n) * \check{N}(n). \tag{10}$$

From above set of equations, we can analyze the effect of $N(n)$ in logarithmic *vs* spectral root cepstrum. As shown in Eq. (4), $\hat{N}(n)$ is additive to $\hat{s}(n)$, its effect in $\hat{r}(n)$ will have limited support. While in Eq. (10), $\check{N}(n)$ is convolved with $\check{s}(n)$. Hence, effect of $\check{N}(n)$ spreads across entire quefrency-domain. This property can be validated from Eq. (8), as it shows that spectral root cepstrum is the linear combination of convolution of logarithmic cepstrum. Hence, additively combined components of logarithmic cepstrum are squeezed in spectral root cepstrum due to convolution. The same analysis is demonstrated in Figure 1. We have chosen 4 synthetic signals, $h(n)$ (Figure 1a), $p(n)$ (Figure 1b), $N(n)$ impulsive (Figure 1c), and $N(n)$ white Gaussian noise (Figure 1d). We depicted two signals, i.e., $s(n) = h(n) * p(n)$ and $r(n) = h(n) * p(n) * N(n)$ in each Figure 1e ($N(n)$ is impulsive) and Figure 1f ($N(n)$ is white Gaussian noise). The effect of $N(n)$ is present across the quefrency domain because of its convolution with $s(n)$. However, in logarithmic cepstrum, the effect of $N(n)$ is concentrated in particular region because of addition of cepstrums (Eq. (4)). Hence, spoofed speech can be discriminated well in spectral root cepstrum than the logarithmic cepstrum.

III. Proposed Feature Extraction

It has been observed that the average auditory nerve firing rate shows an overshoot at the onset of an input signal. In addition, studies shows that the human auditory system appears to focus on the onset of incoming power envelope rather than the falling edge of the same power envelope [24], [25]. Thus, the human auditory system can be modeled by the functional relationship between the onset firing rate of auditory neurons and Sound Pressure Level (SPL). Studies have shown that the given relationship can be approximated by power law nonlinearity [26], [27]. Another advantage of this nonlinearity is that its asymptotic response to the lower amplitude signals approaches to zero rather than negative infinity unlike in MFCC [28]. In addition, it approximates the psychophysical transfer function which relates the physical intensity of sensation to perceived intensity using direct magnitude estimation procedures [29]. Empirically, it is found that various values of gamma gives better speech recognition accuracy for different noise models [30].

In this study, we propose the feature that uses power law non linearity for SSD task. Empirically, it has been observed that it gives better classification accuracy for γ=$-\frac{1}{7}$, as it may detect relevant variability due to $N(n)$ in a spoofed speech. MSRCCs of the time-domain signal $s(n)$ can be obtained by the inverse transformation of spectral energy coefficients raised to certain a exponent γ. Discrete Cosine Transform (DCT) [31] is used to take inverse transformation as it transforms the N real coefficients onto q real independent cepstral coefficients such that $q \ll N$, which extracts the significant information. Mathematical expression of the proposed feature set is given as:

$$MSRCC(q) = \sum_{m=1}^{M} (MFM(m))^\gamma \, cos\left[\frac{q(m-\frac{1}{2})\pi}{M}\right], \tag{11}$$

where the Mel Frequency Magnitude (MFM) spectrum is defined as:

$$MFM(m) = \sum_{k=1}^{K} |S(k)|^2 H_m(k), \tag{12}$$

where $S(k)$ is the k-point DFT of signal $s(n)$, $H_m(k)$ is the triangular weighting-shaped function for the m^{th} Mel scaled bandpass filter.

Information contained in the phase part of STFT is taken into account by PSRCC. As shown in Figure 2, in the development of PSRCC, spectral energy coefficients are replaced by the unwrapped phase. Mathematically,

$$PSRCC(q) = \sum_{m=1}^{M} (MFP(m))^\gamma \, cos\left[\frac{q(m-\frac{1}{2})\pi}{M}\right], \tag{13}$$

where the Mel Frequency Phase (MFP) spectrum is defined

as:

$$MFP(m) = \sum_{k=1}^{K} \angle S(k) H_m(k), \quad (14)$$

where k is the DFT index.

IV. Experimental Results

In this Section, we describe the development of SSD using the proposed feature sets. The experiments are performed on the ASV spoof 2017 challenge database. The full database contains three subsets: training, development (dev), and evaluation (eval) set. The details of ASV spoof 2017 is given in [32], [33].

A. Effect of Gamma Values

We extracted MSRCC features for the frequency range of 6-8 kHz and PSRCC features for the entire auditory frequency range. The *13*-dimensional (D) static MSRCC and PSRCC features along with Δ and $\Delta\Delta$ coefficients are used to get *39*-D feature vector. Total 40 triangular filters along with the Hamming window of 20 ms duration and 50 % overlap are used in both the feature extraction process. The reason behind selecting the 6-8 kHz frequency range for MSRCC feature extraction is that in this range noise magnitude spectrum is more dominant than the speech signal. We perform the experiment for various gamma values. Figure 3, shows the effect of various gamma values on the SSD system. We found empirically that γ=$-\frac{1}{7}$ is the best choice for both the feature sets.

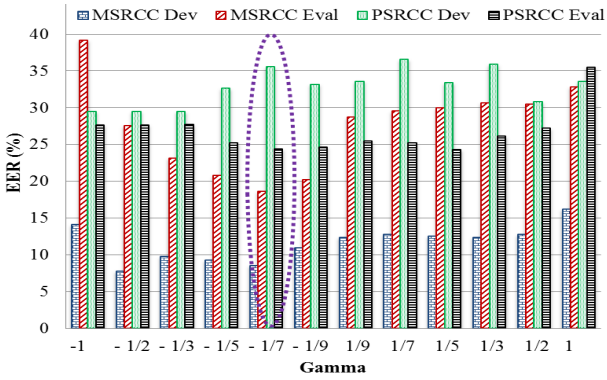

Fig. 3. Bar chart representation showing the effectiveness of gamma values *(Dotted ellipse indicates relatively better result for $\gamma = -1/7$).*

B. Effect of Feature Dimension

We analyze the effect of dimension of MSRCC and PSRCC feature vectors on SSD. In particular, the SSD system with feature dimensions ranging from 21 to 57 *(static+Δ+$\Delta\Delta$)* and 512 Gaussian mixture components is studied. The best choice of gamma from the last experiment is selected (γ=$-\frac{1}{7}$). Figure 4 shows the effect of feature dimension on SSD system. We observe that *39*-D (13-static+13-Δ+13-$\Delta\Delta$) features are sufficient and give relatively best results.

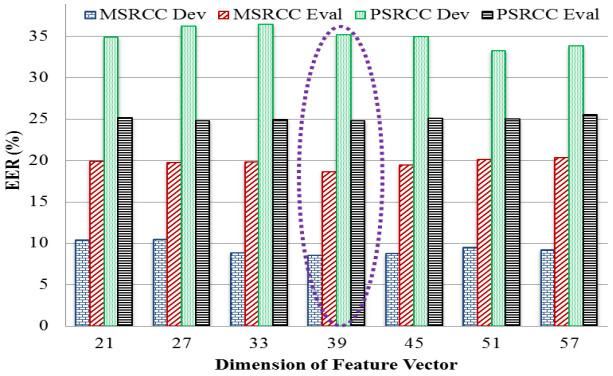

Fig. 4. Bar chart representation showing the effectiveness of feature dimension *(Dotted ellipse indicate relatively better result for 39-D feature vector).*

C. Effect of Number of Gaussian Mixture Components

Furthermore, we examine the effect of the number of Gaussian mixture components. GMM is trained using the training set. We found that the 256 and 512 Gaussian components gave the best results for *39*-D PSRCC and MSRCC feature sets, respectively. The Figure 5 shows the effectiveness of the number of Gaussian components on the SSD system.

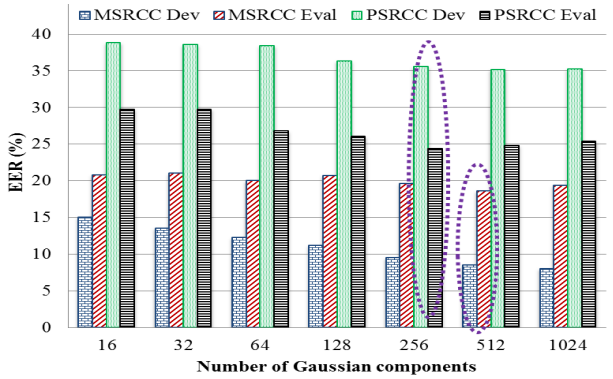

Fig. 5. Bar chart representation showing the effectiveness of number of Gaussian components in GMM *(Dotted ellipse indicate relatively better result for 256 and 512 components for PSRCC and MSRCC respectively).*

D. SSD system using GMM Classifier

In GMM classification, two models are used to represent natural and spoofed speech classes. GMM for each class has 512 and 256 components for MSRCC and PSRCC, respectively. GMM is trained using 30 iterations of Expectation Maximization (EM) algorithm. To explore possible complementary information to that of magnitude part alone, the score-level fusion of MSRCC and PSRCC is investigated. The results of the SSD on development set and evaluation set are presented in Table I.

TABLE I
RESULT (IN % EER) FOR DEV AND EVAL SET

System	Dev	Eval
CQCC-GMM	12.11	29.18
MFCC-GMM	11.21	31.30
MSRCC-GMM	8.53	18.61
PSRCC-GMM	35.53	24.35
MSRCC + PSRCC (GMM)	6.58	**10.65**
MSRCC-CNN	3.05	24.84
PSRCC-CNN	36.21	26.81
MSRCC + PSRCC (CNN)	**2.63**	17.76

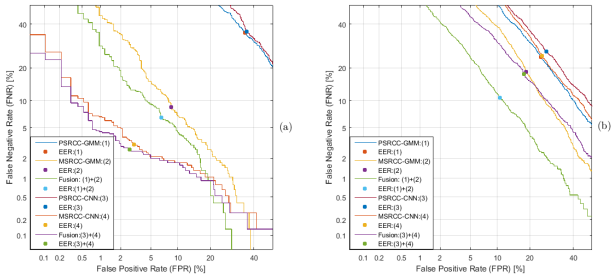

Fig. 6. DET curves for (a) development set and (b) evaluation set of ASV spoof 2017 challange database.

E. SSD System using CNN Classifier

We attempt to use CNN as a classifier to differentiate between the genuine and spoofed speech using MSRCC and PSRCC features. The *13*-D static, Δ and $\Delta\Delta$ are extracted, resulting in *39*-D input feature vector. CNN consists of the three convolutional layers, one max-pooling layer and, three fully-connected layers [16]. All the three convolutional layers have 128 filters and the stride length of 1, whereas max-pooling layer has the kernel of size 1×2 and stride of size 1×2. The network is trained to minimize the cross entropy loss for 70 epochs with an Adam optimization [34] and a learning rate of 0.0001, using an effective batch size of 64. This makes a batch of $(64 \times 39 \times F)$ as an input vector to the CNN and (64×2) generated probabilities using the softmax activation function as an output of the network, with F being a fixed number of frames. Out of 70 epochs, the best model having the least EER on the development and evaluation set are reported in the Table I. Figure 6 shows all the DET curves for development and evaluation set of ASV spoof 2017 database.

V. SUMMARY AND CONCLUSIONS

The spoofing attacks degrade the performance of ASV system and hence, it is necessary to detect them. In this study, we proposed MSRCC and PSRCC feature sets to detect replay spoofed speech. The proposed feature uses the spectral root cepstrum to characterize the natural and replay speech. The individual system is developed and then fused at the score-level using GMM and CNN classifiers. These feature sets contain significant complementary information resulting in improved system performance. The proposed feature set performs better than the ASVspoof 2017 challenge baseline CQCC system. Our future plan is to explore different frequency scale, various filterbank, and different classifiers, such as BLSTM, SVM, etc for the SSD task.

REFERENCES

[1] 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.

[2] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 1678–1681.

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

[4] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *IEEE Annual Summit and Conference in Asia-Pacific Signal and Information Processing Association, (APSIPA)*, Chiang Mai, Thailand, 2014, pp. 1–5.

[5] A. Paul, R. K. Das, R. Sinha, and S. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *International Conference on Signal Processing and Communications (SPCOM)*, Bengaluru, India, 2016, pp. 1–5.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2–6.

[7] R. Font, J. M. Espn, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

[8] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 12–16.

[9] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 22–26.

[10] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 17–21.

[11] M. Witkowski, S. Kacprzak, P. elasko, K. Kowalczyk, and J. Gaka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

[12] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 32–36.

[13] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 82–86.

[14] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 87–91.

[15] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 92–96.

[16] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 97–101.

[17] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 102–106.

[18] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[19] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, India, 2006.

[20] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, pp. 458–465, 1969.

[21] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2014, pp. 1–6.
[22] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun 1968.
[23] J. Lim, "Spectral root homomorphic deconvolution system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.
[24] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 556–565, May 2009.
[25] C. Lemyre, M. Jelinek, and R. Lefebvre, "New approach to voiced onset detection in speech signal and its application for frame error concealment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, March 2008, pp. 4757–4760.
[26] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 106, no. 4, pp. 2040–2050, 1999.
[27] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoustics Research Letters*, vol. 2, no. 3, pp. 91–96, 2001.
[28] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 28–31.
[29] S. S. Stevens, "On the psychophysical law." *Psychological Review*, vol. 64, no. 3, p. 153, 1957.
[30] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, July 2016.
[31] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.
[32] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The REDDOTS data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*. Dresden, Germany: INTERSPEECH, 2015, pp. 2996–3000.
[33] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco *et al.*, "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, USA, 2017, pp. 5395–5399.
[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014, {Last Accessed: March 08, 2018}.