INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Novel Variable Length Teager Energy Separation Based Instantaneous Frequency Features for Replay Detection

Hemant A. Patil, Madhu R. Kamble, Tanvina B. Patel and Meet Soni

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India

{hemant_patil, madhu_kamble, tanvina_bpatel, meet_soni}@daiict.ac.in

Abstract

Replay attacks presents a great risk for Automatic Speaker Verification (ASV) system. In this paper, we propose a novel replay detector based on Variable length Teager Energy Operator-Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (VESA-IFCC) for the ASV spoof 2017 challenge. The key idea here is to exploit the contribution of IF in each subband energy via ESA to capture possible changes in spectral envelope (due to transmission and channel characteristics of replay device) of replayed speech. The IF is computed from narrowband components of speech signal, and DCT is applied in IF to get proposed feature set. We compare the performance of the proposed VESA-IFCC feature set with the features developed for detecting synthetic and voice converted speech. This includes the CQCC, CFCCIF and prosody-based features. On the development set, the proposed VESA-IFCC features when fused at score-level with a variant of CFCCIF and prosody-based features gave the least EER of 0.12 %. On the evaluation set, this combination gave an EER of 18.33 %. However, post-evaluation results of challenge indicate that VESA-IFCC features alone gave the relatively least EER of 14.06 % (i.e., relatively 16.11 % less compared to baseline CQCC) and hence, is a very useful countermeasure to detect replay attacks.

Index Terms: Automatic Speaker Verification, Replay Attack, Variable length Teager Energy Operator, Energy Separation Algorithm, Instantaneous Frequency Cosine Coefficients.

1. Introduction

Automatic Speaker Verification (ASV) system or voice biometrics deals with verifying the claimed identity of a person from his or her voice with help of machines [1]. However, ASV systems are known to be vulnerable to *spoofing* attacks. The various spoofing attacks that exist in the literature include, replay [2], impersonation [3], speech synthesis (SS) [4], voice conversion (VC) [5] and twins [6]. Replay attack deals with playback of pre-recorded speech and it presents a great risk to ASV system [7–9]. An impersonation is an approach where an attacker tries to mimic a genuine target speaker [3, 10]. Likewise, twins can also be considered as imitation based on physiological characteristics [6]. The other machine-generated techniques vulnerable to ASV systems include Text-to-Speech (TTS) synthesis (i.e., generating speech for a given text input) [11] and voice conversion (i.e., manipulating a source speech to sound-like the target speaker through a conversion function) [12–14].

An attempt to develop countermeasures to discriminate genuine speech from the synthetic and voice-converted speech was made in the ASV spoof 2015 challenge [15]. The challenge was based on developing countermeasures for different SS and VC spoofing techniques. In addition to SS and VC speech, the ASV systems are vulnerable to replayed speech [2]. To develop countermeasures for detecting replayed speech, the AVspoof database has been developed [16]. In addition, recently, the ASV spoof 2017 challenge [17, 18] is organized as a follow up as two special sessions on spoofing and countermeasures for ASV held during INTERSPEECH 2013 [19] and 2015 [15]. The ASV spoof 2017 challenge makes use of the recent text-dependent RedDots corpus [20], as well as its replayed version [21]. Thus, the ASV spoof 2017 challenge attempts to interlink the research ideas from spoofing and text-dependent ASV communities. As text-dependent ASV system achieves high verification accuracy with short utterances, it is usually deployed for access control applications such as logic access control in various authentication scenarios.

Replay attack does not require the specific expertise of equipment or tools. The vulnerability of SV to replay attack was evaluated in [7] for the first time. Recent studies concluded that False Alarm Rate (FAR) of ASV system increases due to replay attacks [7–9]. In [22], a replay attack detector was developed in the context of text-dependent SV system. Thereafter, in [9], the use of pitch and Mel Frequency Cepstral Coefficients (MFCC) features were carried to detect cut and paste replayed speech. In [14, 23], the spectral bitmap approach is used to identify live and recorded speech. The use of various cepstral-based features using AVspoof 2015 database has been carried out in [24, 25]. Likewise in [26], the use of long-term spectral statistics is done for spoof detection. In this work, we propose a replay detector based on Variable length Teager Energy Operator-Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (VESA-IFCC). The novelty of the proposed feature set is to exploit the contribution of IF in each subband energy via ESA to capture possible changes in spectral envelope due to transmission and channel characteristics of replay device of replayed speech. The IF is computed from narrowband components of the speech signal, and Discrete Cosine Transform (DCT) is applied on deviations in IF which are referred to as Instantaneous Frequency Cosine Coefficients (IFCCs). We compare the performance of proposed VESA-IFCC features with baseline Constant Q Cepstral Coefficients (CQCC) and other features recently proposed for SS and VC spoofs such as Cochlear Filter Cepstral Coefficients and Instantaneous Frequency (CFCCIF) and prosody-based features.

2. Proposed VESA-IFCC features

2.1. Variable length Energy Separation Algorithm (VESA)

Variable length Teager Energy Operator (VTEO) is the modified version of the traditional TEO method [27]. TEO involves nonlinear operations on the signal, i.e, square of current sample and multiplication of previous and next sample, i.e., $x(n-1)$

http://dx.doi.org/10.21437/Interspeech.2017-1362

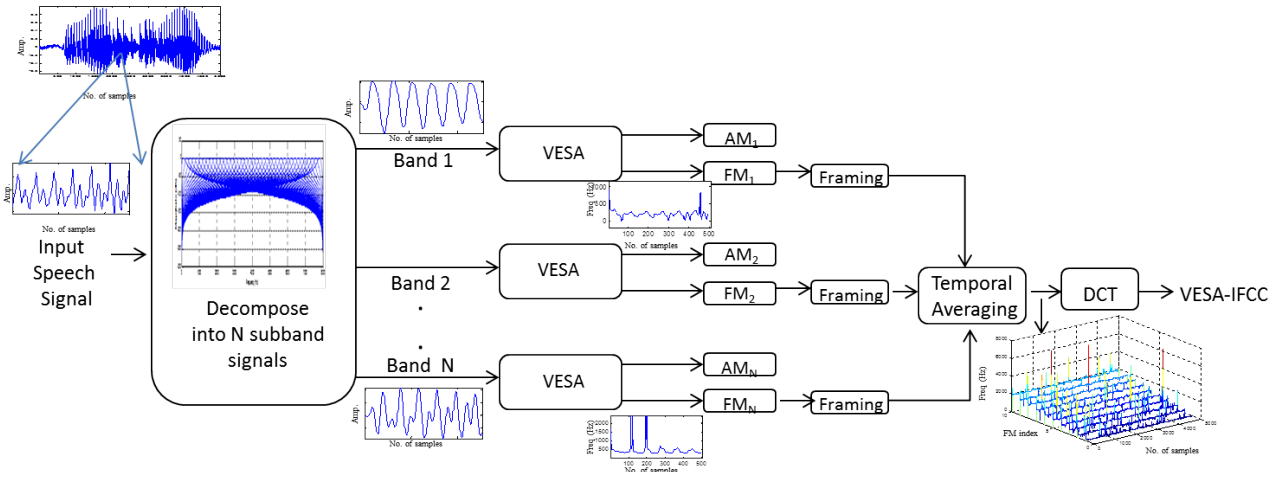

Figure 1: *Schematic diagram to estimate proposed VESA-IFCC feature set. The 3-D plot before DCT corresponds to 5000 samples.*

and $x(n+1)$, respectively. In VTEO algorithm, the number of samples incorporated in energy estimation can be varied up to i past and i future samples, i.e., $x(n-i)$ and $x(n+i)$, instead of only two adjacent samples [28]. VTEO gives flexibility to select these samples to estimate the running estimate of energy required to generate the signal [29]. VTEO gives us a good measure of the energy of the oscillating signal when the sampling rate of the signal is greater than $8i$ times the frequency of oscillation of the signal [28]. VTEO brings out hidden dependencies and dynamics of the signal [28]. For discrete-time signal, $x[n] = Acos(\omega n + \phi)$, its VTEO is given as Eq. (1):

$$E_n = \{\Psi_{DI}\{x(n)\}\} = x^2(n) - x(n-i)x(n+i) \approx i^2A^2\omega^2, \quad (1)$$

where $i^2A^2\omega^2$ is instantaneous estimate of signals energy multiplied by i^2 and referred as VTEO for the dependency index (DI), i, which is expected to give running estimate of signal's energy [29, 30]. To estimate the individual contribution of amplitude $a[n]$ and frequency $\omega[n]$ of signal, Maragos et.al [31], [32] developed an *Energy Separation Algorithm (ESA)* that uses nonlinear energy operator (i.e., in TEO framework) to track the instantaneous energy of the source generating the AM-FM signal and separate it into its amplitude and frequency components. The ESA was developed to demodulate a speech signal into amplitude envelope (AE) and IF. According to Kaiser, energy in a speech is a function of both amplitude and frequency [33]. However, ESA is applied to single speech resonance, while the speech signal itself is multi-component, being the sum of several resonances. Hence, there is a need to isolate resonances by bandpass filtering. In this paper, we propose to exploit VTEO to track the modulation energy and estimate the instantaneous amplitude and frequency of AM-FM signal and refer to it as VESA. The IF $\omega[n]$ and AE $a[n]$ at any time the instant of the AM-FM modulated signal $x[n]$ is given by:

$$a_i[n] \approx \frac{2\Psi_{DI}\{x[n]\}}{\sqrt{\Psi_{DI}\{x[n+1]-x[n-1]\}}}, \quad (2)$$

$$\omega_i[n] \approx arcsin\sqrt{\frac{\Psi_{DI}\{x[n+1]-x[n-1]\}}{4\Psi_{DI}\{x[n]\}}}. \quad (3)$$

Eq. (2) and Eq. (3) reduces to original ESA algorithm when DI=1, for $\Psi_{DI}\{x(n)\} = TEO(x(n))$. The frequency estimation part assumes that $0 < \omega_i[n] < \frac{\pi}{2}$ because the computer implementation of $arcsin(u)$ function assumes that $|u| < \frac{\pi}{2}$. Thus, discrete ESA can be used to estimate IF $< 1/4$ of sampling frequency of signal [31]. The IF is modeled as the superposition of slow and fast-varying components. The slow-varying component models the average formant frequency values and the fast-varying component models frequency variations around the formant frequency.

2.2. Proposed VESA-IFCC Feature Set

Fig. 1 shows the block diagram of proposed VESA-IFCC feature set. Here, the input speech signal is first split into N frequency subband signals. The ESA is applied using VTEO with various dependency index (DI) ($i = 1$ to 10) onto each N bandpass (subband) filtered signals to obtain corresponding AEs and IFs. Furthermore, we have discarded the AE and taken only IF and computed for each of the narrowband components in order to emphasize the spectral envelope of genuine *vs.* replayed speech. The IF are segmented into overlapping short (segmental) frames of 20 ms duration, shifted by 10 ms, and the temporal average is computed to obtain N-dimensional IFCs for every frame. The redundancy among IFCs is exploited to obtain a low-dimensional representation by employing DCT that has energy compaction property and thus, retaining first few DCT coefficients that are referred to as Instantaneous Frequency Cosine Coefficients (IFCC). The IFCC along with their delta and double delta features were also appended resulting in higher-dimensional feature set denoted as VESA-IFCC. Algorithm 1 shows the procedure for extracting VESA-IFCC features.

2.3. Spectrographic Analysis

The Butterworth filter provides a maximally flat response (i.e., the first $2n-1$ derivatives for the power spectrum w.r.t. frequency are zero and hence, has no ripples) in the passband. Butterworth filter has a nonlinear phase that can be approximated as linear over smaller frequency regions. Butterworth filterbank is used with filters placed according to linear scale. In earlier studies [34], *linearly* spaced equi-bandwidth filters are more suitable for IF computation than Mel-spaced varying bandwidth filters. In case of Mel filterbank, the bandwidth increases at high frequencies, making the computation of IF less reliable and hence, we have used linear frequency scale [35]. For a given 16 kHz sampling frequency, we have an available bandwidth of 7800 Hz that is divided into 40 equi-spaced fre-

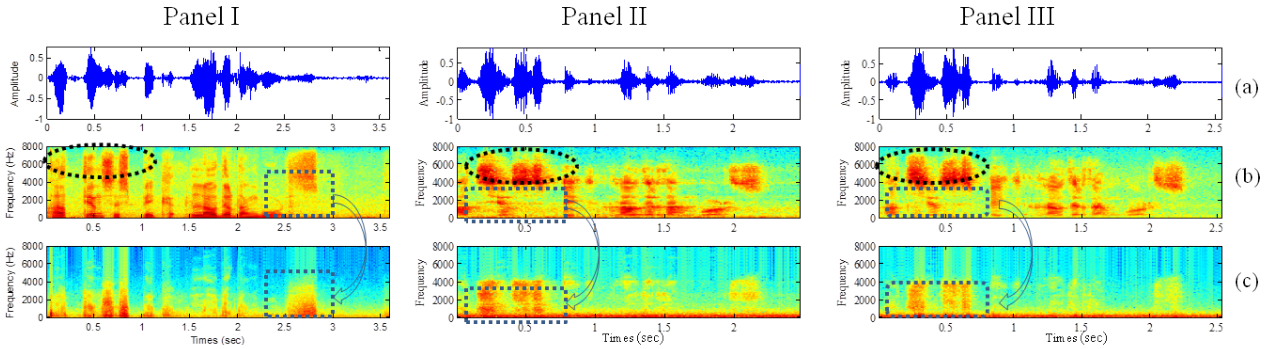

Figure 2: *Spectrographic analysis for utterance, "Actions speak louder than words" (a) speech signal (b) corresponding spectrogram (c) spectral energy density of 40 subband filtered signals (i.e., N=40 in Fig. 1). Panel I for natural speech, Panel II for replay speech with recording done in balcony and Panel III for replay speech with recording done in the bedroom.*

quency regions of width $(f_H - f_L)/40$ Hz. The phase response around each $(f_H - f_L)/40$ Hz width is mostly found to be linear (as observed in author's recent study reported in [36]). Fig. 2 shows the spectrographic analysis of natural (Panel I), replayed speech recorded in the balcony (Panel II) and replayed speech recorded in the bedroom (Panel III) speech signals. Fig. 2(a) shows the time-domain speech signal and its corresponding spectrogram is shown in Fig. 2(b) whereas the spectrogram obtained after 40 subbands Butterworth filtered signals is shown in Fig. 2(c). It can be observed that the lower frequency regions corresponding to lower spectral amplitude (for the lower vocal tract resonances, also referred to as formants) are emphasized more in spectrogram obtained after Butterworth filtered signals for genuine and replayed speech whereas lower frequency regions are absent in the short-time Fourier transform (STFT) spectrogram. It can be observed that spectral energy is reduced significantly in higher frequency regions (especially 5-8 kHz) and overall spectral smearing/blunting of spectral resolution (possibly due to convolution of impulse response due to transmission loss and characterization of replay device indicating microphone, speaker and recording environment [37]) for replayed speech.

Algorithm 1 The VESA-IFCC feature extraction from speech

1: $x(n)$= speech signal.
2: Consider an N channel filterbank with linearly spaced Butterworth filters in time-domain.
3: **for** i=1 to N **do**
4: Perform narrowband filtering of $x(n)$ through i^{th} filter; $x_i(n)$.
5: Compute VTEO from $x_i(n)$ as in Eq. (1)
6: Compute ESA and extract IF $\omega_i(n)$ as in Eq. (3).
7: **end for**
8: Segment $\omega_i(n)$, $i = 1, 2.....N$ into short-time frames of duration as 20 ms, shifted by 10 ms.
9: Average IF for each frame to obtain N-dimensional IFCs.
10: Apply DCT on VESA-IFCs and retain first few coefficients to get VESA-IFCCs.
11: Append VESA-IFCCs with their first and second-order derivatives.

3. Experimental Setup

Following state-of-the-art feature sets that were explored for the detection of SS and VC spoof are used here for detection of replayed speech.

Constant Q Cepstral Coefficients (CQCC): CQCC features are extracted with the constant Q transform (CQT) that employs a variable time-frequency resolution [37]. The CQCC features are extracted with $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8 kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \approx 15Hz$. The number of bins per octave B is set to 96. Features extracted with 30 DCT static coefficients (with log-energy), resulting in total 90-D feature vector [38].

Cochlear Filter Cepstral Coefficients and Instantaneous Frequency (CFCCIF): The CFCCIF features were used by the authors in the first ASVspoof 2015 challenge that makes use of envelope of the output of each cochlear filter and its IF for spoof detection [39, 42]. For the present task of replay detection both natural and replay speech is from the human speaker, hence, the derivative operation is eliminated from original CFCCIF method. The CFCCIF features are estimated with 40 filterbanks and using a frameshift of 25 ms and 50 % overlap. Features extracted with 12 DCT static coefficients (without log-energy) resulting in the 36-D feature vector.

Prosody Features: This includes the use of F_0 contour and strength of excitation (SoE) estimated at the glottal closure instants (GCIs) in the voiced regions. In addition, F_0 and of SoE is estimated from the speech signal through zero frequency (ZF) filtering method [40]. In [41] the dynamics of exploring F_0 contour and SoE's as source-based features were used for spoof detection. For the present problem, dynamic variations up to the 5^{th} order are used as an 18-D feature vector.

3.1. ASV Spoof 2017 Database and Model Training

The challenge is based on the recent text-dependent RedDots corpus and its replayed version [20,21]. The former serving as a source of genuine recordings and the latter as a source of replay spoof recordings. The details of the database are given in [17]. We have used Gaussian Mixture Model (GMM) with 512 mixtures for modeling the two class classifier in which the classes correspond to genuine and replayed utterances. Final scores are represented in terms of Log-Likelihood Ratio (LLR). The decision of the test speech being genuine or replay is based on

the LLR. To obtain the complementary information of CQCC, CFCCIF, Prosody, MFCC and VESA-IFCC features, we use their score-level fusion as in our other studies [41, 42]. The performance is measured by computing the Equal Error Rate (EER) as in [17].

4. Experimental Results

4.1. Selection of Dependency Index and Feature Dimension

Table 1 shows the % EER for various DI's in VTEO. It was observed from Table 1 that for DI=9 classification between genuine and replayed speech performs relatively better than other DI's. Thus, the performance of replay detector is optimized w.r.t DI (as in our earlier work [29]). The proposed VESA-IFCC are estimated with 40 filters in filterbank with a framesize of 20 ms, shifted of 10 ms and 40 coefficients (with log-energy), which are appended by their first and second-order derivatives, resulting in 120 -D feature vector. From Table 2, it is observed that the performance of proposed feature set when extracted with 40 number of coefficients gives relatively least EER as compared to the different dimension of feature vector.

Table 1: *Effect of DI in VESA-IFCC on the development set*

DI	1	2	3	4	5	6	7	8	9	10
EER	7.65	6.61	6.65	6.63	6.99	7.41	8.57	6.46	**4.61**	7.17

Table 2: *Effect of feature dimension (FD) on the development set in terms of % EER for DI=9*

FD	39	60	90	120
EER	8.02	7.25	7.59	**4.61**

4.2. Results on ASV spoof 2017 Database

The results of the development and evaluation set for the individual performance of CQCC (baseline system), CFCCIF, prosody and VESA-IFCC features are shown in Table 3. It is observed that on development set, the baseline system, i.e., CQCC gave 11.06 % EER. The prosody-based features gave an individual EER of 29.40 % and CFCCIF alone gave 6.08 %. The proposed VESA-IFCC features gave the relatively best individual performance of 4.61 %. On the development set, it is observed that the relatively best performing system is VESA-IFCC + CFCCIF + prosody with a score-level fusion factor with weight of 0.69, 0.23, 0.03, respectively, giving the best EER of 0.127 %. This same fusion factor is used for the evaluation set and we obtained EER of 18.33 %, which was our primary submission for the ASV spoof 2017 challenge.

Table 3: *The result in % EER on development and evaluation set with GMM classifier. * Primary submission, ** Post evaluation*

Feature set	Development	Evaluation
CQCC (baseline)	11.06	30.17
A: CFCCIF	6.8	34.49**
Prosody	29.40	31.40**
B: VESA-IFCC	4.61	**14.06****
VESA-IFCC+MFCC	1.47	17.93**
VESA-IFCC+CQCC	2.08	15.35**
A+B+Prosody	**0.1263**	**18.33***

The DET curves of the same features (as in Table 3) is shown in Fig. 3. It is observed that with the fusion of the CFCCIF, prosody features, the performance of the VESA-IFCC features improves to 0.127 as compared to its original EER of 4.61 %. Two baseline results for CQCC, namely, EER of 30.17 %

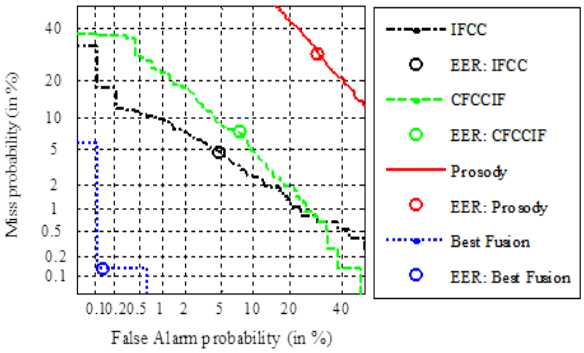

Figure 3: *The individual DET curves for VESA-IFCC, CFCCIF, prosody and the best fusion factor on the development set.*

(trained data) and 24.65 % (train+development data, i.e., pooled data) are provided in the ASV spoof 2017 challenge overview paper [18]. Table 4 shows the results using the pooled dataset. For the VESA-IFCC+CFCCIF+Prosody combination, the EER obtained using pooled data is 23.68 %, which degrades as compared to the EER obtained without pooling. Interestingly, the use of pooled data for VESA-IFCC results in better performance on development set with decrease in EER from 4.61 % to 3.42 %. However, for evaluation set it increases from 14.06 % to 15.50 % indicating further challenge and scope for this task.

Table 4: *Results with pooled dataset*

Feature set	EER
CQCC (baseline)	24.65
VESA-IFCC	**15.50**
VESA-IFCC+CFCCIF+Prosody	23.68

5. Summary and Conclusions

In this study, we proposed novel VESA-IFCC features to capture characteristics of natural *vs.* replayed speech. Proposed feature set exploit contribution of individual IFs in each subband energies via proposed VTEO-based ESA algorithm. Spectrographic analysis demonstrated effectiveness of proposed approach to discriminate replayed speech from natural speech w.r.t difference in spectral energy density (in high frequency regions) and spectral smearing due to replay device. For ASV spoof 2017 challenge task, proposed feature set performed relatively better than baseline CQCC and recently proposed CFCCIF. Moreover, our post evaluation results indicated the superior performance of VESA-IFCC than CFCCIF and CQCC. Our future work will be directed towards exploring other filterbank, number of subbands, etc in proposed features.

6. 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] 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)*, 2014, pp. 1–6.

[3] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 145–148.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric

speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[5] Y. Stylianou, "Voice transformation: A survey," in *IEEE ICASSP*, 2009, pp. 3585–3588.

[6] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[7] J. Lindberg, M. Blomberg *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques." in *EUROSPEECH*, vol. 99, 1999, pp. 1211–1214.

[8] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.

[9] Villalba, Jesús and Lleida, Eduardo, "Detecting replay attacks from far-field recordings on speaker verification systems," in *European Workshop on Biometrics and Identity Management*. Springer, 2011, pp. 274–285.

[10] Y. W. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the yoho speaker verification corpus," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2005, pp. 15–21.

[11] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system." in *INTERSPEECH*, 2001, pp. 759–762.

[12] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates." in *INTERSPEECH*, 2007, pp. 2053–2056.

[13] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *IEEE ICASSP*, 2012, pp. 4401–4404.

[14] 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 Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, 2014, pp. 1–5.

[15] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[16] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[17] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2017.

[18] Kinnunen, Tomi and Evans, Nicholas and Yamagishi, Junichi and Lee, Kong Aik and Sahidullah, Md and Todisco, Massimiliano and Delgado, Héctor, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection." submitted in *INTERSPEECH*, 2017.

[19] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *INTERSPEECH*, 2013, pp. 925–929.

[20] K.-A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The reddots data collection for speaker recognition." in *INTERSPEECH*, 2015, pp. 2996–3000.

[21] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. A. L. Thomsen, A. K. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE ICASSP*, 2017.

[22] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *IEEE ICASSP*, 2010, pp. 1678–1681.

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

[24] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *INTERSPEECH*, 2016.

[25] D. Paul, M. Sahidullah, and G. Saha, "Generalization of spoofing countermeasures: A case study with asvspoof 2015 and btas 2016 corpora," in *IEEE ICASSP*, New Orleans, USA, 2017, pp. 2047–2051.

[26] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–6.

[27] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *IEEE ICASSP*, Albuquerque, New Mexico, USA, 1990, pp. 381–384.

[28] V. Tomar and H. A. Patil, "On the development of variable length Teager energy operator (VTEO)." in *INTERSPEECH*, 2008, pp. 1056–1059.

[29] H. A. Patil and K. K. Parhi, "Novel variable length Teager energy based features for person recognition from their hum," in *IEEE ICASSP*, 2010, pp. 4526–4529.

[30] J. Choi and T. Kim, "Neural action potential detector using multi-resolution teo," *Electronics Letters*, vol. 38, no. 12, pp. 541–543, 2002.

[31] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[32] Maragos, Petros and Kaiser, James F and Quatieri, Thomas F, "On separating amplitude from frequency modulations using energy operators," in *IEEE ICASSP*, vol. 2, San Francisco, California, USA, 1992, pp. 1–4.

[33] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America (JASA)*, vol. 99, no. 6, pp. 3795–3806, 1996.

[34] P. R. Reddy, K. Vijayan, and K. S. R. Murty, "Analysis of features from analytic representation of speech using MP-ABX measures,," in *Proc. INTERSPEECH*, Dresden, Germany, 2015.

[35] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[36] P. B. Bachhav, H. A. Patil, and T. B. Patel, "A novel filtering based approach for epoch extraction," in *IEEE ICASSP*, 2015, pp. 4784–4788.

[37] Todisco, Massimiliano and Delgado, Héctor and Evans, Nicholas, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, 2017.

[38] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[39] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural *vs.* spoofed speech,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2062–2066.

[40] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[41] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F0) and strength of excitation (SoE) for spoofed speech detection," in *IEEE ICASSP*, 2016, pp. 5105–5109.

[42] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *accepted in IEEE Journal of Selected Topics in Signal Processing*, Dec. 2016.