Interspeech 2018
2-6 September 2018, Hyderabad

Novel Variable Length Energy Separation Algorithm using Instantaneous Amplitude Features For Replay Detection

Madhu R. Kamble and Hemant A. Patil

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

{madhu_kamble, hemant_patil}@daiict.ac.in

Abstract

Voice-based speaker authentication or Automatic Speaker Verification (ASV) system is now becoming practical reality after several decades of research. However, still this technology is very much susceptible to various spoofing attacks. Among various spoofing attacks, replay is the most challenging attack. In this paper, we propose a novel feature set based on our recently introduced Variable length Energy Separation Algorithm (VESA) during INTERSPEECH 2017. The key idea of this paper is to capture the Instantaneous Amplitude (IA) obtained from the instantaneous energy fluctuations. The replay speech is affected by acoustic environment and distortions of intermediate device. Thus, the noise added in replayed speech is important to detect. The Amplitude Modulations (AM) are more susceptible to noise and multipath interferences that may result due to replay mechanism. The experiments are performed on various dependency index (DI) and lower EER of 6.12 % and 11.94 % is found on dev and eval set, respectively, of ASV Spoof 2017 Challenge database. Furthermore, we compare our results with CQCC, LFCC, MFCC, and VESA-IFCC feature sets. The score-level fusion VESA-IFCC and proposed feature set further reduced the EER to 0.19 % and 7.11 % on dev and eval set, respectively.

Index Terms: Automatic Speaker Verification, Spoofing, Energy Separation Algorithm, Instantaneous Amplitude.

1. Introduction

Voice of a speaker is one of the key attribute that can be used in the voice biometric or Automatic Speaker Verification (ASV) system that can be used for security purpose such as banking transactions, secure personalized data, etc. However, one of the major threat to the ASV system is that they are vulnerable to various kinds of spoofing attacks, namely, speech synthesis [1], voice conversion [2], replay [3], twins and impersonation [4]. The large availability and the widespread usage of the mobile/smart gadgets, recording devices, it is easy and simple to record the speaker's voice, without having the prior information which makes the replay spoof attacks simple to produce and execute. The challenging task, in replay attack is to detect acoustical characteristics of replayed speech, as there is a imperceptible difference of the speech signal between the natural and the replayed speech. The speech signal recorded with the playback device contains the convolutional and additive distortions from the intermediate devices [5]. It is therefore very important to develop countermeasures that can detect spoofing attacks.

In the recent years, several countermeasures were proposed targeting either text-dependent or text-independent ASV systems [6]. In addition, countermeasures for replay attacks on text-independent ASV systems were evaluated on AV spoof dataset [7, 8]. The 2^{nd} ASV Spoof 2017 Challenge focused on text-dependent replay attacks [9]. The organizers have used RedDots corpus and its replayed version to generate the replay attacks under different playback, recording and environmental conditions [10, 11]. A baseline system is provided by the organizers of the Challenge, the system uses Constant-Q Cepstral Coefficients (CQCC) with a Gaussian Mixture Model (GMM) as a classifier [9, 12]. The performance of the baseline system is not good enough to be used as a countermeasure in the ASV system and hence, there was a need to develop the standalone countermeasure. Several countermeasures were approached during ASV Spoof 2017 Challenge. The different acoustic features were studied in [13] and found Subband Spectral Centroid Magnitude Coefficients (SCMCs) followed by feature normalization performed better for replay detection. The Instantaneous Frequency (IF)-based feature sets were explored in [14, 15]. The study in [16] used high resolution temporal features known as Single Frequency Filter (SFF). For replay detection, high frequency region is found to be more useful [17, 18]. Few approaches used deep learning along with feature normalization [19–21].

Replay detection task is to identify the recorded speech signal from a live speaker or an intermediate (recording + playback) device. Because of convolutional distortion introduced by the intermediate device, the replay speech contains lower damping compared to the live speech signal [5]. In replay detection, the feature extraction is the important part that identifies the characteristics of the intermediate device.

In our earlier study, we used VESA along with estimating the Instantaneous Frequency (IF) [14]. In this paper, we extended our earlier work by exploiting the Instantaneous Amplitude (IA) information and replaced the Butterworth filterbank with linearly-scaled Gabor filterbank. The IA obtained from the Amplitude Modulation (AM) component of a speech signal is severely affected by the noise and multipath interference because of presence of the replay mechanism. The novelty of this paper lies in exploiting this degradation in AM component as a signature of replayed speech than its natural counterpart.

2. AM-FM Demodulation

The Teager Energy Operator (TEO) is a nonlinear operator that tracks the instantaneous energy of a signal [22]. The Energy Separation Algorithm (ESA) is a demodulation technique used along with TEO [23], [24]. The demodulation techniques is even used in the auditory cortex [25]. The amplitude and frequency components of a signal can be obtained after ESA algorithm. The TEO for a discrete signal $\{\Psi_d\}$ is defined as the squared product of amplitude and frequency and is given

10.21437/Interspeech.2018-1687

by [23]:

$$\Psi_d\{x(n)\} = x^2(n) - x(n-1)x(n+1) \approx A^2\Omega^2, \quad (1)$$

The TEO of a speech signal, $x[n] = a[n]cos(\phi[n])$ is applied to obtain AM-FM component and it can approximately estimate the squared product of Instantaneous Amplitude (IA) ($a_i[n]$) and Instantaneous Frequency (IF) ($\Omega_i[n]$) of signals. The TEO for the i^{th} subband speech signal is defined as:

$$\Psi_d\left(a_i[n]cos\left(\int_0^n \Omega_i[m]dm + \theta\right)\right) \approx a_i^2[n]\Omega_i^2[n]. \quad (2)$$

The TEO operates with 3 samples for a given instant of time, i.e., it uses present, past and future sample values. The generalized TEO was by replacing 1 with a constant arbitrary integer j, i.e., varying the samples of the past and future signal, i.e., $x(n-j)$ and $x(n+j)$, instead of only two adjacent samples [26, 27]. This constant arbitrary integer is called as *lag parameter* [28]. This *lag parameter* can be varied having value greater than 1 and thus, we define the generalized TEO as Variable Length version of TEO, i.e., VTEO and is given by [29]:

$$\Psi_{DI}\{x(n)\} = x^2(n) - x(n-j)x(n+j) \approx j^2A^2\Omega^2, \quad (3)$$

The advantage of VTEO over TEO lies in the superior localization and tracking instantaneous fluctuations (if any) of the energy at a given instant of time and also it brings out the *hidden* dependencies and dynamics of the signal w.r.t. distantly located speech samples than only immediate adjacent samples. The VTEO have a good measure of the energy when the sampling rate is greater than $8i$ times the frequency of the oscillation of signal [29]. This instantaneous energy can be decomposed using Variable length Energy Separation Algorithm (VESA) and obtain the Instantaneous Amplitude and Instantaneous Frequency (IA-IF) of a speech signal. The IA and IF of a VTEO signal is given by [23]:

$$a_i[n] \approx \frac{2\Psi_{DI}\{x[n]\}}{\sqrt{\Psi_{DI}\{x[n+1]-x[n-1]\}}}, \quad (4)$$

$$\Omega_i[n] \approx arcsin\sqrt{\frac{\Psi_{DI}\{x[n+1]-x[n-1]\}}{4\Psi_{DI}\{x[n]\}}}. \quad (5)$$

3. Proposed Feature Extraction

The block diagram of AM-based feature extraction using ESA of a VTEO signal is shown in Figure 1. Here, the input speech signal is passed through a pre-emphasis filter to emphasize the higher frequency regions [30]. The TEO works on a mono-component speech signal and as speech signal is a multicomponent signal, we need to pass the signal through a bandpass filter to obtain a subband filtered signals at various cut-off frequencies. Here, we have used a linearly-scaled Gabor filterbank to obtain the subband filtered signals. The AM-FM modulation features corresponding to the i^{th} subband filtered signal are extracted from instantaneous frequency $f_i(t)$ and amplitude envelope $a_i(t)$, where i=1,2,...., L, and L is the number of subband filtered signals [31], i.e.,

$$r_i(t) = a_i(t)cos\left(2\pi\int_0^t f_i(\tau)d\tau\right), \quad (6)$$

where $r_i(t) \approx s(t) * g_i(t)$, $s(t)$ is the speech signal and $g_i(t)$ is the impulse response of the i^{th} Gabor filter. We have used Gabor filter $g(t)$ as it is compact and smooth (i.e., $g(t)\ \epsilon\ C^\infty$ which is function space of infinitely differentiable functions), and hence, it has *optimal* joint time-frequency resolution (since Fourier transform (FT) of Gaussian is Gaussian) [32]. The impulse response of the Gabor filter is given by [22]:

$$g(t) = exp(-b^2t^2)cos(\omega_c t), \quad (7)$$

where ω_c is the center frequency of the filter and b is a parameter for controlling the bandwidth of a filter. The subband filtered signal obtained from Gabor filter is then passed through demodulation algorithm, i.e., VESA. We computed IA and IF for each i^{th} subband signals from the VESA algorithm, and further discarded IF signal and focused on IA part. These IA's were averaged over a short-time window of 20 ms with 10 ms shift to obtain L-dimensional Instantaneous Amplitude Coefficients (VESA-IAC) for each frame. The low-dimensional feature vectors are obtained by applying Discrete Cosine Transform (DCT) on VESA-IAC (to de-correlate the features while training the models) and will be denoted as Instantaneous Amplitude Cosine Coefficients, i.e., VESA-IACC. We have used Cepstral Mean Normalization (CMN) as a post-processing technique, to overcome the channel mismatch/distortion between the training and testing feature vectors [33, 34].

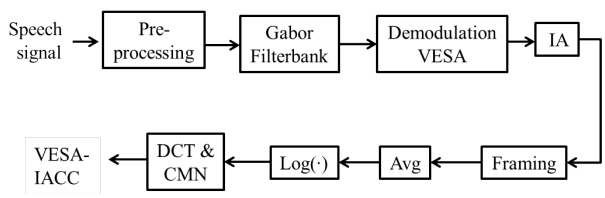


Figure 1: *Schematic block diagram of proposed VESA-IACC feature set.*

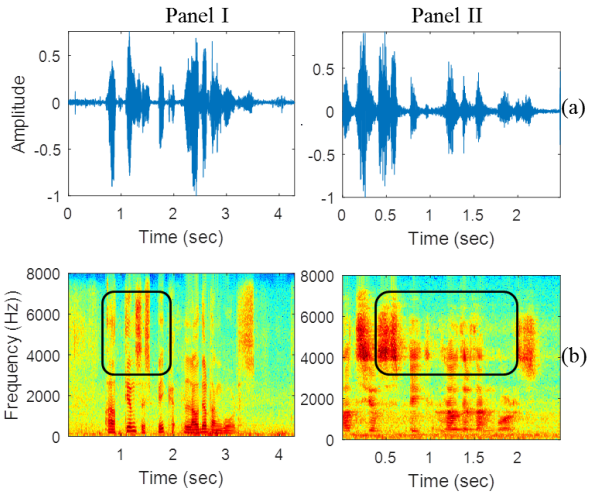


Figure 2: *Spectral energy densities of Panel I (natural) and Panel II (replay) speech signal. (a) time-domain speech signal and (b) spectral energy density obtained from 40 linearly-spaced Gabor filtered subband signal.*

The spectral energy density obtained from 40 linearly-scaled Gabor filterbank is shown for natural (Panel I) and replayed (Panel II) speech signal as shown in Figure 2. The formants and harmonics of the natural speech signal are clearly

observed in Panel I of Figure 2(b). On the other hand, for replayed signal, the formants and harmonics are *blunted* in the lower frequency regions. While the high frequency regions have relatively higher energy (indicated by rectangular box in Figure 2(b)) because of the distortion involved in replayed signal due to the intermediate devices and various acoustical conditions.

4. Experimental Setup

We have used ASV Spoof 2017 Challenge version 1 database that consists of replay spoof attacks with text-dependent system [9]. The database is prepared from the RedDots corpus and its replayed version [10, 11]. The detailed description of the database is given in [9].

4.1. Feature Parameters

The proposed feature set, i.e., VESA-IACC are extracted from 40 linearly-scaled Gabor filterbank and further a window of 20 ms with a shift of 10 ms was used and averaged each frame block. Finally, 40-Discrete Cosine Transform (DCT) static coefficients were appended to their first and second-order derivatives resulting to the *120*-dimensional (D) feature vector. Along with VESA-IACC feature set, we have also used VESA-IFCC feature set from our earlier study [14]. The VESA-IFCC feature set were extracted from 40 linearly-scaled Butterworth filterbank with 40-DCT static coefficients resulting in total *120*-D feature vector. We have compared our results of proposed feature set with three standard feature sets in the spoofing detection task, namely, CQCC, LFCC, and MFCC. The feature parameters used for CQCC is 30-DCT static coefficients along with Δ and $\Delta\Delta$. For LFCC feature set we extracted *120*-D feature vector (40 static+Δ+$\Delta\Delta$) and for MFCC, we used *39*-D feature vector (13 static+Δ+$\Delta\Delta$). The Gaussian Mixture Model (GMM) is used as a classifier to classify the speech signal being natural or replayed signal. The number of Gaussian components used in GMM to train the models from the training dataset is 512. The decision of the final scores of speech signal being natural or replayed speech is decided by the Log-Likelihood Ratio (LLR) and is given by:

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \quad (8)$$

where $P(X|H_0)$, and $P(X|H_1)$ are the likelihood scores of natural and replay trials. The score-level fusion of two feature sets is given by:

$$LLK_{fused} = \alpha LLK_{feature1} + (1-\alpha)LLK_{feature2}, \quad (9)$$

where $LLK_{feature1}$ is a log-likelihood score of CQCC, LFCC, and MFCC, whereas $LLK_{feature2}$ proposed feature set. The fusion parameter (α) lies between $0 < \alpha < 1$ to decide the weight of scores.

5. Experimental Results

5.1. Results for Various Dependency Index (DI)

5.1.1. Results on Development Set

The results with varying the DI from 1 to 4 on development (dev) set of the proposed feature set VESA-IACC are shown in Table 1. The results of our recently proposed VESA-IFCC are also shown Table 1 [14]. The VESA-IACC feature set obtained an EER of 6.12 % with DI=1, whereas with VESA-IFCC the EER is 6.61 % on DI=2. Since VESA-IACC and VESA-IFCC capture distinct information of amplitude and frequency to explore possible complementary information captured by them, their score-level fusion is done Thus, to explore this individual information of both the feature sets, we have fused these features for all the four DI's. From Table 1, it can be observed that after score-level fusion of VESA-IACC and VESA-IFCC for each DI, the EER is reduced significantly than that for individual feature sets indicating that these two feature sets indeed capture complementary information individual feature sets. The best lower EER was obtained with the fusion of features at DI=4 resulting in the reduced EER of 0.19 % from 7.18 % (VESA-IACC) and 6.63 % (VESA-IFCC) feature set clearly demonstrating the potential of idea of exploring DI in TEO [29].

Table 1: *Results on Dev Dataset of Proposed Feature Set on Various Dependency Index (DI)*

DI	VESA-IACC	VESA-IFCC [14]	Fusion
1	**6.12**	7.65	1.72
2	7.44	**6.61**	0.33
3	7.83	6.65	0.26
4	7.18	6.63	**0.19**

5.1.2. Results on Evaluation Set

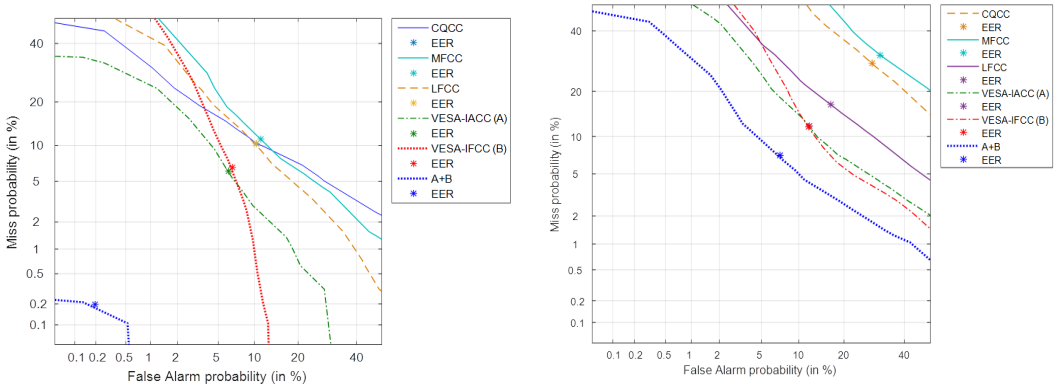
Similarly, we computed our proposed features on evaluation (eval) set as done on dev set. We varied the DI from 1 to 4 for both VESA-IACC and VESA-IFCC feature sets. The lower EER was obtained on VESA-IACC is with DI=2 of 11.94 %, while with VESA-IFCC feature set, we got 11.79 % with DI=4. The best lower EER was obtained with the score-level fusion of VESA-IACC and VESA-IFCC feature set reducing the EER to 7.11 % on DI=4 from the individual EER of 12.27 % (VESA-IACC) and 11.79 % (VESA-IFCC) feature set.

Table 2: *Results on Eval Dataset of Proposed Feature Set on Various Dependency Index (DI)*

DI	VESA-IACC	VESA-IFCC	Fusion
1	12.08	16.16	9.11
2	**11.94**	13.46	7.56
3	12.09	12.34	7.15
4	12.27	**11.79**	**7.11**

5.2. Results of Score-Level Fusion

To explore the possible complementary information present in other feature sets, namely, CQCC, LFCC, and MFCC, we have used the scores of those features and fused them at score-level with VESA-IACC and VESA-IFCC as shown in Table 3. The score-level fusion is performed for every DI on both dev and eval datasets. The score-level fusion with CQCC indeed helped to reduce the EER for each DI. On dev set, the lower EER of 3.99 % was obtained when fused with DI=1 in VESA-IACC feature set, while with DI=4 the fusion of VESA-IFCC and CQCC gave lower EER of 3.28 %. The next fusion was done with LFCC feature set, the score-level fusion of VESA-IACC and LFCC did not reduce the EER, whereas the fusion with VESA-IFCC reduced the EER to 0.38 % with DI=4. The VESA-IACC feature set was extracted with linear scale and LFCC also uses the linear scale and hence, possibly the score-level fusion did not reduce the EER as both are magnitude spectrum-based features and thus, may not carry much complementary information. For most of the DI's, the lower EER obtained was the same as that of the VESA-IACC feature



(a) (b)

Figure 3: *DET curves on dev and eval datasets. (a) the individual DET curve on dev set of CQCC, MFCC, LFCC, VESA-IACC (A), VESA-IFCC (B) and score-level fusion A+B and (b) similar DET curves on eval set.*

Table 3: *Results on Score-Level Fusion of CQCC, LFCC, and MFCC with Various Dependency Index (DI) on Dev and Eval Dataset*

		Dependency Index (DI)							
		VESA-IACC				**VESA-IFCC**			
	DI	1	2	3	4	1	2	3	4
Dev	CQCC	**3.99**	4.27	4.40	4.39	5.32	3.75	3.64	**3.28**
	LFCC	**6.12**	7.44	7.38	7.18	2.23	0.74	0.56	**0.38**
	MFCC	**4.06**	4.36	4.39	4.31	3.21	**1.42**	2.26	1.58
Eval	CQCC	11.18	**11.13**	11.28	11.49	16.16	13.46	12.34	**11.79**
	LFCC	12.08	**11.94**	12.09	12.27	10.32	8.44	7.84	**7.93**
	MFCC	12.08	**11.94**	12.09	12.27	16.16	13.46	12.34	**11.79**

set. On the other hand, the fusion of VESA-IFCC and LFCC also uses linear scale, however, they carry the complementary information of magnitude and phase spectrum because of which it reduced the significantly. Finally, we fused our feature sets with MFCC obtaining an EER of 4.06 % with VESA-IACC for (DI=1) and 1.42 % for (DI=2) with VESA-IFCC features. Similarly, the score-level fusion was performed on the eval dataset. There was a reduction in the EER when fused with VESA-IACC (for DI=1) and CQCC resulting in 11.13 %, whereas the fusion of VESA-IACC (for DI=2) with LFCC and MFCC obtained reduced EER of 11.94 % as shown in Table 3. On the other hand, the score-level fusion of VESA-IFCC with CQCC and MFCC did not reduce the EER for all the DI's. While the score-level fusion of VESA-IFCC (with DI=4) and LFCC reduce the EER from the individual system to 7.93 %. Table 4 shows the final results of our proposed feature set. The organizers of ASV Spoof 2017 Challenge provided CQCC feature set with GMM classifier as the baseline system. In this paper, we have considered CQCC and LFCC as two distinct baselines systems. The proposed feature set was extracted with linearly-spaced Gabor filterbank and thus, to compare our proposed results with a linearly-scaled feature set, we consider LFCC as the second baseline. At last, we used one of the well known MFCC feature set to compare our results. The EER of all the feature sets, namely, CQCC, LFCC, and MFCC are high on both dev and eval dataset. The EER for CQCC (baseline system) gave an EER of 10.21 % and 28.48 % on dev and eval set, respectively. The VESA-IACC and VESA-IFCC feature sets individually performed better than the CQCC, LFCC, and MFCC feature sets. The best results obtained with the score-level fusion of VESA-IACC and VESA-IFCC resulting in the lower EER of 0.19 % on dev set and 7.11 % on eval set.

The performance is also shown by the DET curve in Figure 3(a) for dev set and Figure 3(b) for eval set for CQCC, MFCC, LFCC, VESA-IACC and VESA-IFCC feature set along with score-level fusion of VESA-IACC and VESA-IFCC. On dev and eval set, score-level fusion of VESA-IACC and VESA-IFCC are clearly distinct at *all* the operating points of the DET curve and have a significantly lower false alarm and miss probabilities in the DET curve compared to the CQCC, LFCC, and MFCC feature set.

Table 4: *Final Results on Dev and Eval Dataset*

Feature Set	**Dev**	**Eval**
CQCC (Baseline)	10.21	28.48
LFCC	10.58	16.62
MFCC	11.21	31.30
A:VESA-IACC	6.12	11.94
B:VESA-IFCC	6.61	11.79
A+B	**0.19**	**7.11**

+: score-level fusion

6. Summary and Conclusions

In this study, we investigate the advantage of VESA over ESA with varying the DI to capture the *hidden* dependencies and dynamics for replay SSD task. The VESA algorithm has the superior localization and tracking instantaneous energy properties that makes to estimate accurately the IA and IF signals. We found that the estimated VESA-IACC and VESA-IFCC feature sets perform better than the baseline systems. The reduced EER clearly demonstrates the potential idea of exploring DI in ESA. Furthermore, the score-level fusion of both VESA-IACC and VESA-IFCC component captures the possible significant complementary information of each other and reduced the EER further than the individual systems. In future, we plan to explore the effect of different filterbanks and investigate the characteristics of the intermediate device.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Y. Stylianou, "Voice transformation: A survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Taipei, Taiwan, China, 2009, pp. 3585–3588.

[3] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH, Lyon, France*, 2013, pp. 940–944.

[4] 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*, Hong Kong, 2004, pp. 145–148.

[5] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in ASV systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, 2017, pp. 51–55.

[6] 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.

[7] 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)*, Bangalore, India, 2016, pp. 1–5.

[8] P. Korshunov, S. Marcel, H. Muckenhirn, A. Gonçalves, A. S. Mello, R. V. Violato, F. O. Simões, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Niagara Falls, New York, USA, 2016, pp. 1–6.

[9] 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. 1–6.

[10] 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 *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, Louisiana, USA, 2017, pp. 5395–5399.

[11] 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*, Dresden, Germany, 2015, pp. 2996–3000.

[12] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[13] R. Font, J. M. Espín, 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.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 1991, pp. 421–424.

[23] P. Maragos, T. F. Quatieri, and J. Kaiser, "On separating amplitude from frequency modulations using energy operators," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, San Francisco, California, USA, 1992, pp. 1–4.

[24] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[25] A. Dimitrijevic, M. S. John, P. van Roon, and T. W. Picton, "Human auditory steady-state responses to tones independently modulated in both frequency and amplitude," *Ear and hearing*, vol. 22, no. 2, pp. 100–111, 2001.

[26] P. Maragos and A. Potamianos, "Higher order differential energy operators," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 152–154, 1995.

[27] J. Choi and T. Kim, "Neural action potential detector using multiresolution TEO," *Electronics Letters*, vol. 38, no. 12, pp. 541–543, 2002.

[28] W. Lin, C. Hamilton, and P. Chitrapu, "A generalization to the Teager-Kaiser energy function and application to resolving two closely-spaced tones," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Detroit, Michigan, USA, 1995, pp. 1637–1640.

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

[30] L. Deng and D. O'Shaughnessy, *Speech Processing – A Dynamic and Optimization-Oriented Approach*. 1^{st} Edition, Marcel Dekker Inc., June 2003.

[31] D. Dimitriadis and E. Bocchieri, "Use of micro-modulation features in large vocabulary continuous speech recognition tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1348–1357, 2015.

[32] S. Mallat, *A Wavelet Tour of Signal Processing. 2^{nd} Edition*. Academic press, 1999.

[33] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 1, Hong Kong, China, 2003, pp. I–656–659–I.

[34] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 1, Phoenix, Arizona, USA, 1999, pp. 325–328.