Speech Demodulation-based Techniques for Replay and Presentation Attack Detection

Madhu R. Kamble*, Pulikonda Aditya Krishna Sai†, Maddala V. Siva Krishna†, Ankur T. Patil*,
Rajul Acharya*, Hemant A. Patil*
* Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India
E-mail: {madhu_kamble, ankur_patil, rajul_acharya, hemant_patil}@daiict.ac.in
† Indian Institute of Information Technology (IIIT), Vadodara, Gujarat, India
E-mail: {201551013, 201551045}@iiitvadodara.ac.in

***Abstract*—Spoofing is one of the threats that bypass the voice biometrics and gains the access to the system. In particular, Automatic Speaker Verification (ASV) system is vulnerable to various kinds of spoofing attacks. This paper is an extension of our earlier work, the combination of different speech demodulation techniques, such as Hilbert Transform (HT), Energy Separation Algorithm (ESA), and its Variable length version (VESA) is investigated for replay Spoof Speech Detection (SSD) task. In particular, the feature sets are developed using Instantaneous Amplitude and Instantaneous Frequency (IA-IF) components of narrowband filtered speech signals obtained from linearly-spaced Gabor filterbank. We observed relative effectiveness of these demodulation techniques on two spoof speech databases, i.e., BTAS 2016 and ASVspoof 2017 version 2.0 challenge database that focus on the presentation and replay attacks, respectively. The results obtained from different demodulation techniques gave comparable results on both databases showing small variations in % Equal Error Rate (EER). For VESA, we found that with Dependency Index (DI) = *2* gave relatively better performance compared to the other DI on both the databases for SSD task. All the demodulation technique-based feature sets gave lower % EER than their baseline system for both the databases.**

Index Terms: Spoof, Presentation Attack, Hilbert Transform, Teager Energy Operator.

I. Introduction

Automatic Speaker Verification (ASV) system grants access to the system by verifying the claimed identity of speaker. However, due to recent advances in technology, the claimed identity could be generated by malicious means or other resources. The resources include different ways of speech generation also known as the spoofing attacks, i.e., voice conversion (VC), speech synthesis (SS), replay, twins, and impersonation [1], [2], [3]. The present advanced technology demands for more robust ASV systems to the spoofing attacks to sustain the technology race. The awareness of the Spoof Speech Detection (SSD) task and its countermeasures was widely spread throughout the globe with the help of ASVspoof 2013 special session [1]. The successor ASVspoof 2015 challenge has focused on the countermeasures for the machine generated speech, such as VC and SS [4]. Recording and playback of the target speaker's speech sample is the easiest way that any fraud person can prefer to break the ASV system [5]. This attack is known as the replay attack and it poses the highest threat due to its easy implementation. The Biometrics: Theory, Applications, and Systems (BTAS) 2016 Speaker Anti-Spoofing competition and ASVspoof 2017 challenge have focused on the replay detection.

The BTAS 2016 competition used the AVspoof database that consists of replayed speech signals of natural and machine generated (i.e., SS and VC) signals using intermediate devices, such as high quality speakers, laptop speakers, and mobile phones [6]. Whereas, the ASVspoof 2017 challenge database contains the replay spoof signals that are recorded using different recording and playback devices in different uncontrolled real acoustic environments [7]. Several countermeasures were designed and submitted by participants in both challenges. Some of the countermeasures for BTAS 2016 used Mel Frequency Cepstral Coefficients (MFCC), Inverse Mel Frequency Cepstral Coefficients (IMFCC), and normalized perceptual linear predictive features as front-end feature sets. The organizers of the BTAS 2016 competition provided a baseline using simple spectrogram-based ratios as features and logistic regression as classifier [8].

For ASVspoof 2017 challenge, Constant-Q Cepstral Coefficients (CQCC) with Gaussian Mixture Models (GMM) classifier is provided as baseline by the challenge organizers [9], [10]. Some of the countermeasures includes the normalization techniques and various acoustic features [11]. Instantaneous Frequency (IF)-based features were explored in [12], [13]. Also, high-resolution temporal-based features such as, Single Frequency Filtering (SFF) [14], high frequency band selection of CQCC [15], modulation dynamic features, and temporal modulation features [16], [17], were used for designing the countermeasures. Neural Network (NN)-based classifiers, such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Bi-directional Long Short Term Memory (BLSTM) [18], [19], [14] were also explored in the challenge.

In this paper, we are exploring our earlier speech demodulation techniques- based on Instantaneous Amplitude and Instantaneous Frequency (IA-IF) components for two different databases, i.e., ASVspoof 2017 challenge version 2.0 (v2.0) (for replay classification), and BTAS 2016 database (for presentation attack detection). In particular, Hilbert Transform (HT), Energy Separation Algorithm (ESA) and its Variable

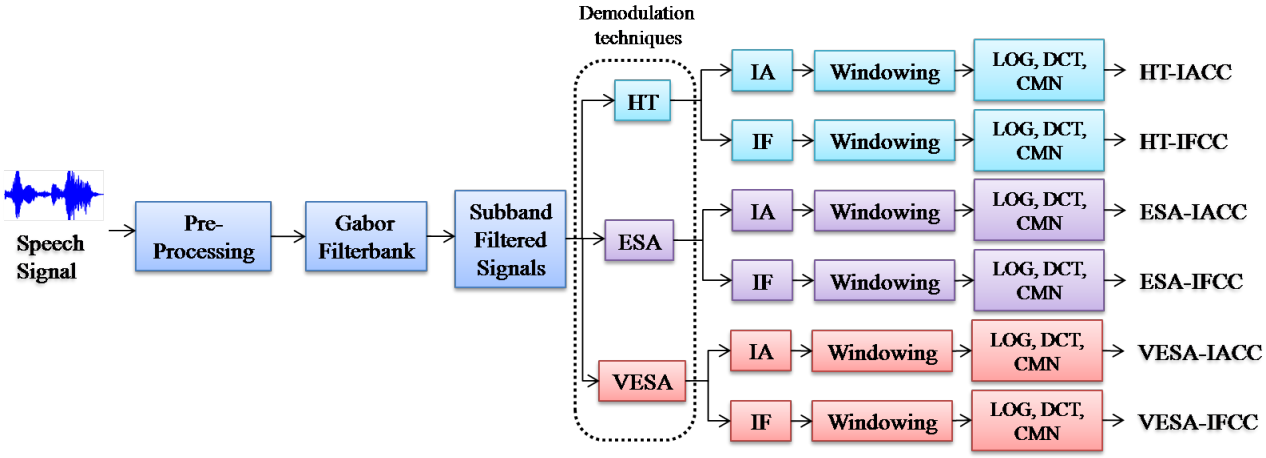

Fig. 1. Block diagram for feature extraction of IA and IF-based features using HT, ESA, and VESA.

length version (VESA) are studied to compute their IA and IF components obtained from narrowband signals for SSD task.

II. Speech Demodulation-based Features

The Amplitude and Frequency Modulations features (AM-FM) computed using three different demodulation techniques, such as HT, ESA, and VESA are discussed in this section.

A. Hilbert Transform (HT):

The Hilbert transform estimates amplitude envelope and frequency function of a speech signal [20], [21]. Let $s_a(t)$ be the analytic signal corresponding to the real signal, $s(t)$, then $s_a(t)$ is given by:

$$s_a(t) = s(t) + j\hat{s}(t), \tag{1}$$

where quadrature signal $\hat{s}(t)$ is the Hilbert transform of $s(t)$ and $\phi(t)$ represents the phase. The Instantaneous Amplitude (IA), $a_h(t)$, and Instantaneous Frequency (IF), $\phi'_h(t)$, are derived from the analytic signal as:

$$IA_{HT} = a_h(t) = \sqrt{s^2(t) + \hat{s}^2(t)}, \tag{2}$$

$$IF_{HT} = \phi'_h(t) = \frac{d}{dt}(\phi(t)). \tag{3}$$

B. Energy Separation Algorithm (ESA):

For a discrete-time monocomponent signal, $x[n]$, the Teager Energy Operator (TEO), $\Psi_d\{\cdot\}$, is defined as [22], [23]:

$$E_n = \Psi_d\{x[n]\} = x^2[n] - x[n-1]x[n+1] \approx A^2\Omega^2, \tag{4}$$

where E_n gives the running estimate of signal's energy, A is amplitude and Ω is frequency (in radians). The speech signal can be considered as the combination of several monocomponent signals and TEO works on narrowband signal. Hence, bandpass filtering is necessary to apply on the input speech signal to compute 'N' number of subband filtered signals. The Teager energy obtained from the subband signals are further separated into IA ($a_i[n]$) and IF ($\Omega_i[n]$) components for the i^{th} subband filtered signal, using Energy Separation Algorithm (ESA) and it is given as [24], [25], [26]:

$$IA_{ESA} = a_i[n] \approx \frac{2\Psi_d\{x_i[n]\}}{\sqrt{\Psi_d\{x_i[n+1] - x_i[n-1]\}}}, \tag{5}$$

$$IF_{ESA} = \Omega_i[n] \approx arcsin\sqrt{\frac{\Psi_d\{x_i[n+1] - x_i[n-1]\}}{4\Psi_d\{x_i[n]\}}}. \tag{6}$$

C. Variable length Energy Separation Algorithm (VESA):

The TEO operates with 3 samples for a given instant of time, i.e., $x(n)$, $x(n-1)$, and $x(n+1)$. The generalized TEO replaces 1 with a constant arbitrary integer k, i.e., varying the samples of the past and future signal, i.e., $x(n-k)$ and $x(n+k)$ [27], [28]. This constant arbitrary integer is known as *lag parameter* (also known as Dependency Index (DI)) and it can be varied from the value greater than 1 and thus, named as Variable Teager Energy Operator (VTEO) which is given as: [13], [29], [30], [31]

$$\Psi_{DI}\{x(n)\} = x^2(n) - x(n-k)x(n+k) \approx k^2A^2\Omega^2. \tag{7}$$

Similar to above ESA technique, we can compute the IA and IF from VESA by replacing the VTEO in the place of TEO, i.e., as:

$$IA_{VESA} = a_i[n] \approx \frac{2\Psi_{DI}\{x_i[n]\}}{\sqrt{\Psi_{DI}\{x_i[n+1] - x_i[n-1]\}}}, \tag{8}$$

$$IF_{VESA} = \Omega_i[n] \approx arcsin\sqrt{\frac{\Psi_{DI}\{x_i[n+1] - x_i[n-1]\}}{4\Psi_{DI}\{x_i[n]\}}}. \tag{9}$$

The block diagram of speech demodulation technique-based features are shown in Figure 1. The IA and IF component-based feature sets proposed in the earlier studies are reported in [13], [26], [31], [32]. Initially, signal is passed through the pre-emphasis filter, and then passed through the filterbank to obtain N number of subband signals [23], [24], [33]. We used linearly-spaced Gabor filterbank to have almost equal

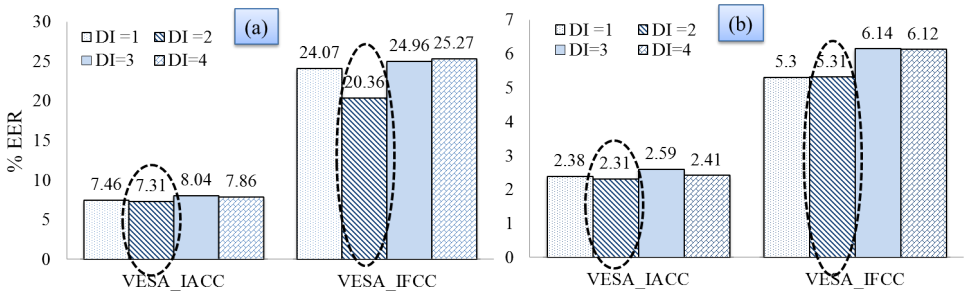

Fig. 2. Varying DI from 1 to 4 on development set of (a) ASVspoof 2017 challenge v2.0, and (b) BTAS 2016 competition database.

bandwidth to cover the entire frequency range [26], [31], [34]. Furthermore, these subband filtered signals are given as input to the HT, ESA, and VESA block to compute corresponding IA and IF components. These individual IA and IF components are passed through the frame blocking and averaging using a short window length of 20 ms with a shift of 10 ms followed by logarithm operation to compress the data. The Discrete Cosine Transform (DCT) and Cepstral Mean Normalization (CMN) technique is then applied for energy compaction and retained first few DCT coefficients to obtain HT, ESA, and VESA-based IA and IF Cepstral Coefficients i.e., (IACC and IFCC), followed by their Δ and $\Delta\Delta$ feature vector to obtain higher-dimensional feature vector.

The spectral energy density obtained from all the three speech demodulation techniques are shown in Figure 3 for a time-domain speech signal (a). The corresponding spectral energy for HT is shown in Figure 3(b), for ESA it is shown in Figure 3(c), and for VESA it is shown in Figure 3(d). The highlighted dotted box in the Figure 3 shows the spectral differences for all the three different techniques. With VESA-based spectral energy it can be observed that the high resolution for the harmonics and frequency bands in the lower frequency region is obtained.

D. Databases Used

In this section, we provide the details of databases used, the evaluation metrics, feature parameters along with classifier.

1) ASVspoof 2017 Challenge v2.0 Database: The ASVspoof 2017 challenge v2.0 database mainly depends on the RedDots corpus, and its replayed speech [9], [35]. The organizers of ASVspoof 2017 challenge provided a baseline system using CQCC as features and GMM as classifier. The detailed statistics of the database is given in [9], [36], [37].

2) BTAS 2016 Database: The BTAS 2016 database is based on the publicly available AVspoof database [6]. We have used the same database that was provided in the BTAS 2016 competition . The organizers of the BTAS 2016 competition provided a baseline system using simple spectrogram-based ratios as features and logistic regression as classifier. The detailed statistics of the database is given in [8].

E. Evaluation Metrics

The evaluation metrics considered in this paper are according to the protocol used in the BTAS 2016 speaker anti-spoofing challenge. The results on the development data are reported in terms of Equal Error Rate (% EER) and on the test data in terms of Half Total Error Rate (% HTER).

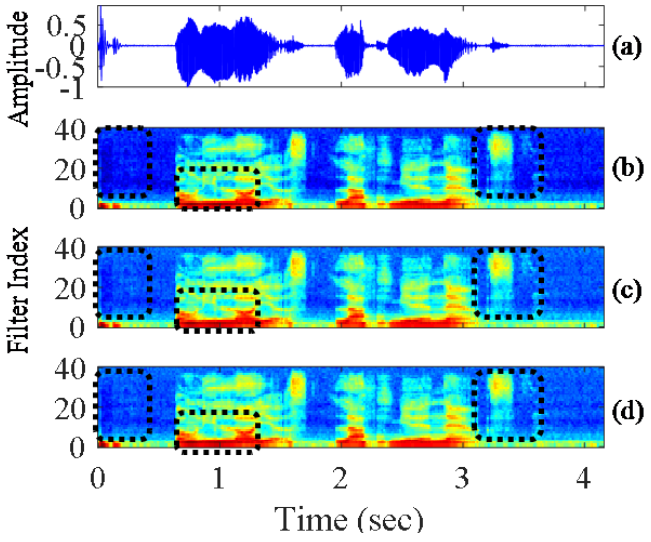

Fig. 3. (a) Time-domain speech signal, and its corresponding spectral energy densities using (b) HT, (c) ESA, and (d) VESA with DI=2.

The evaluation of the replay attacks systems was based on the *false rejection rate* (FRR) and *false acceptance rate* (FAR), that in turn depend on a threshold θ. We use the development set to determine threshold θ_{dev}. The evaluation performance of the system is then computed as the HTER given as :

$$\theta_{dev} = \arg\min_{\theta} \frac{\text{FAR}_{dev}(\theta) + \text{FRR}_{dev}(\theta)}{2}, \quad (10)$$

$$\text{HTER}_{eval}(\theta) = \frac{\text{FAR}_{eval}(\theta_{dev}) + \text{FRR}_{eval}(\theta_{dev})}{2}. \quad (11)$$

F. Features and Classifier

For the experimentation, we have used IACC and IFCC each of which extracted using HT, ESA and VESA-based approaches. The features are extracted using 40 linearly-spaced Gabor filterbank with f_{min} = 10 Hz, and f_{max} = 8000 Hz. For each subband filtered signals, we obtained *40* - dimensional (D) static features appended with their Δ and $\Delta\Delta$ coefficients resulting in *120* - dimensional (D) feature

vector which are used as features for our SSD system.

For the classification of natural *vs.* replayed speech, we have used the GMM as the classfier [38]. For experiments performed on the ASVspoof 2017 challenge v2.0 database, 512 Gaussian mixtures are used whereas 64 Gaussian mixtures are used for BTAS 2016 challenge due to computational complexities resulting from huge data provided in training set for BTAS 2016 challenge.

III. Experimental Results

A. VESA-based Results with Varying DI

The results with varying the lag parameter also called as Dependency Index (DI) from 1 to 4 on development set for VESA-IACC and VESA-IFCC feature set on both databases (i.e., (a) ASVspoof 2017 challenge v2.0 and (b) BTAS 2016 competition are shown in Figure 2). It can be observed that both IA and IF-based feature sets gave lower % EER at DI=2 on both databases. On ASVspoof 2017 challenge v2.0 database, the % EER varies from 7.31 % to 8.04 % for IA-based features whereas for IF-based features it varies from 20.36 % to 25.27 %. On the other hand, for BTAS 2016 database, the variation is from 2.31 % to 2.59 % and from 5.3 % to 6.14 % for IA and IF-based features, respectively. Hence, for further set of experiments reported in this paper VESA-based features are extracted using DI=2.

B. Results on Development Set

Results on all the speech demodulation techniques for both ASVspoof 2017 challenge v2.0 and BTAS 2016 database are reported in Table I and Table II, respectively. It can be observed that on development set, HT-based features gave lower % EER, whereas, for evaluation set VESA-based features gave better performance that other two demodulation techniques. However, on BTAS database the results varies for all the speech demodulation techniques with very less differences in % EER. The advantage of VESA over ESA lies in its superior localization and approximation to track the instantaneous fluctuations (if any) of the energy at a given instant of time. The VESA brings out the *hidden* dependencies and dynamics of the signal w.r.t. distantly located speech samples than only immediate adjacent samples.

TABLE I
Results of IACC and IFCC feature sets using HT, ESA, and VESA on ASVspoof 2017 challenge v2.0 database (in % EER)

	IACC		IFCC	
	Dev	**Eval**	**Dev**	**Eval**
HT	**7.16**	12.58	**18.86**	30.18
ESA	7.99	13.45	24.07	19.87
VESA (DI=2)	7.31	**12.57**	20.36	**19.10**

To obtain the possible complimentary information between two feature sets, we used score-level fusion of two feature set obtained from same demodulation techniques. For example, the IA and IF components extracted from HT-based method are fused together to obtain the reduced % EER and give high performance. It can be observed from Table III that with score-level fusion, on both the databases we reduced the % EER from its individual % EER. We compared our ASVspoof 2017 challenge v2.0 results with the baseline system of the same database, i.e., CQCC feature set. The baseline system gave % EER of 12.81 % and 19.04 % on development and evaluation set, respectively. The best % EER obtained on development set is with HT-based method giving an % EER of 5.91 %, and on evaluation set the lower % EER is obtained with VESA-based technique resulting in 11.45 %

TABLE II
Results of IACC and IFCC feature sets using HT, ESA, and VESA on BTAS 2016 database (in % EER)

	IACC		IFCC	
	Dev	**Eval**	**Dev**	**Eval**
HT	**2.26**	**3.96**	5.26	**7.46**
ESA	2.36	4.31	**5.08**	9.23
VESA (DI=2)	2.31	4.73	5.31	9.13

TABLE III
Results of score-level fusion of IACC and IFCC feature sets using HT, ESA, and VESA on ASVspoof 2017 v2.0 and BTAS 2016 database (in % EER)

	ASVspoof 2017 v2.0		BTAS 2016	
	Dev	**Eval**	**Dev**	**Eval**
CQCC (Baseline)	12.81	19.04	–	–
HT	**5.91**	12.13	**2.26**	**3.93**
ESA	7.72	12.17	2.36	4.31
VESA (DI=2)	6.99	**11.45**	5.31	4.73

Table IV shows the performance on evaluation set in % HTER on BTAS 2016 database and compared our results with the baseline system. The baseline system gave an % HTER of 6.87 % and the best performance of our speech demodulation technique obtain an % HTER of 3.17 % with IA component obtained from HT-based technique. The performance is also

TABLE IV
% HTER for Eval set of BTAS 2016

System Used	% HTER
Baseline	6.87
HT-IACC	**3.17**
HT-IFCC	6.74
ESA-IACC	3.64
ESA-IFCC	7.59
VESA-IACC (DI=2)	4.06
VESA-IFCC (DI=2)	7.14

shown by the Detection Error Trade-off (DET) curve in Figure 4 on (a) development and (b) evaluation set of ASVspoof 2017 challenge v2.0 database. The DET curves are shown only for the score-level fusion of IA and IF components of individual demodulation techniques i.e., HT, ESA, and VESA, respectively. It can be observed from the DET curves for development set that HT-based technique gave lower % EER with less miss probability and false alarm rate. However, for the evaluation set, the HT technique did not perform well and with VESA method it performed better. This fluctuations in the performance brings out more generalized countermeasure

for SSD task. Note: we have not shown the DET curves for BTAS 2016 database results because the score-level fusion did not reduce the % EER from the individual IA-based results.

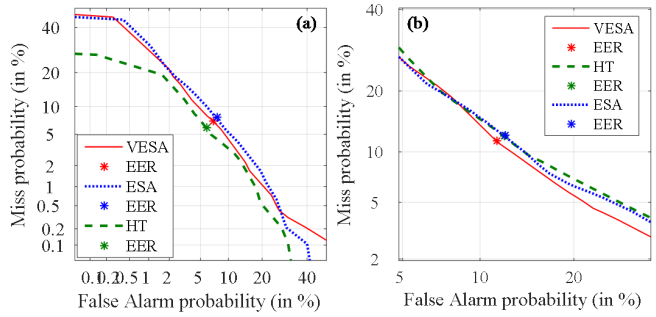

Fig. 4. DET curves of score-level fusion of IACC and IFCC on (a) dev and (b) eval set of ASVspoof 2017 challenge v2.0 database.

IV. Summary and Conclusions

In this paper, we analyzed and studied different speech demodulation techniques, namely, Hilbert Transform (HT), Energy Separation Algorithm (ESA), and Variable length version of ES, i.e., VESA. The speech demodulation-based features are used for spoof speech detection task, to classify the replay and presentation attack from natural speech. We investigated the advantage of VESA over HT and ESA by varying the Dependency Index (DI) to capture the *hidden* dependencies and dynamics. The features obtained from different demodulation techniques gave better results than the baseline system of both ASVspoof 2017 challenge v2.0 and BTAS 2016 database. Furthermore, the score-level fusion is performed on both IA and IF components to capture the possible significant complementary information of each other and reduced the % EER further than the individual systems.

References

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

[2] N. Evans, J. Yamagishi, and T. Kinnunen, "Spoofing and countermeasures for speaker verification: A need for standard corpora, protocols and metrics," *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, pp. 2013–05, 2013.

[3] 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.

[4] 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.

[5] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2014, pp. 1–6.

[6] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE 7^{th} International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[7] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, and M. Todisco, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[8] P. Korshunov *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition," in *IEEE 8^{th} International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–6.

[9] T. Kinnunen, M. Sahidullah *et al.*, "The ASVspoof 2017 Challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1–6.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] M. Witkowski *et al.*, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

[16] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation dynamic features for the detection of replay attacks," in *INTERSPEECH*, 2018, pp. 691–695.

[17] H. Sailor, M. Kamble, and H. A. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," in *INTERSPEECH*, 2018, pp. 666–670.

[18] 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.

[19] 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.

[20] L. Cohen, *Time-Frequency Analysis. 1^{st} Editon*. Prentice Hall PTR Englewood Cliffs, NJ:, 1995, vol. 778.

[21] R. Sharma, L. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner, and S. Prasanna, "Empirical mode decomposition for adaptive AM-FM analysis of speech: A review," *Speech Communication*, vol. 88, pp. 39–64, 2017.

[22] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, USA, 1990, pp. 381–384.

[23] 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.

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

[25] 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.

[26] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 641–645.

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

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

[29] 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.

[30] V. Tomar and H. A. Patil, "On the development of variable length teager

energy operator (VTEO)." in *INTERSPEECH*. Brisbane, Australia: Citeseer, 2008, pp. 1056–1059.

[31] M. R. Kamble and H. A. Patil, "Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 646–650.

[32] M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," in *IEEE European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 106–110.

[33] D. Dimitrios, M. Petros, and P. Alexandros, "Auditory Teager energy cepstrum coefficients for robust speech recognition." in *INTERSPEECH*, Lisboa, Portugal, 2005, pp. 3013–3016.

[34] M. R. Kamble and H. A. Patil, "Novel amplitude weighted frequency modulation features for replay spoof detection," *in ISCSLP*, Taipei, Taiwan 2018.

[35] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. Pan 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.

[36] 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.

[37] H. Delgado, M. Todisco *et al.*, "ASVspoof 2017 version 2.0: Meta data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, pp. 296–303.

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