Significance of Teager Energy Operator Phase for Replay Spoof Detection

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

***Abstract*—The increased use of voice biometrics for various security applications, motivated authors to investigate different countermeasures for the hazard of spoofing attacks, where the attacker tries to imitate the genuine speaker. The replay is the most accessible spoofing attack. Past studies have ignored phase information for various speech processing applications. In this paper, we explore the excitation source-like feature set, namely, Teager Energy Operator (TEO) phase and its significance in the replay spoof detection task. This feature set is further fused at score-level with magnitude spectrum-based features, such as Constant Q Cepstral Coefficients (CQCC), Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). The improvement in the results show that the TEO phase feature set contains the complementary information to the magnitude spectrum-based features. The experiments are performed on the ASV Spoof 2017 Challenge database. The systems are implemented with Gaussian Mixture Model (GMM) as a classifier. Our best system using TEO phase achieves the Equal Error Rate (EER) of 6.57 % and 15.39 % on the development and evaluation set, respectively.**

I. Introduction

Due to significant advancement in speech technology, Automatic Speaker Verification (ASV) exists to be reliable biometric solution for the various applications [1]. For practical applications, the ASV system need to be robust against variations, such as transmission channel and microphone, intersession, acoustic noise, speaker aging, etc. This robustness makes ASV system to be vulnerable to various spoofing attacks as it tries to nullify these effects and make replayed speech much similar to the natural speech. Hence, we would like the system to be secure against spoofing attacks. ASV systems are susceptible to five types of spoofing attacks, namely, impersonation [2], [3], Voice Conversion (VC) [4], [5], Speech Synthesis (SS) [6], [7], replay [8], [9], and twins [10], [11]. Among them, the replay is the most accessible attack as it does not require any special computer skills or complex algorithms as in case of VC and SS, also it posses a greater risk to the ASV system [1]. In the replay attack, an attacker tries to access the speaker's identity by original speaker's pre-recorded speech [12]. In 2017, the second ASV spoof challenge was organized for the detection of replay attacks [13]. The replay spoof detection task is to decide whether the given input speech is genuine or replay speech signal. The replay speech can be modeled as convolution of natural speech with impulse response of recording device, impulse response of playback device, impulse response of recording environment and impulse response of playback environment [9]. Hence, the detection difficulty increases with a high quality intermediate devices, clean recording and playback environment, because in such cases the replay speech is close to the natural speech.

The first approach for replay spoof detection was reported in [14]. In this study, the authors discussed score-normalization approach for replay attack detection for text-dependent ASV. Authors of [15], [16] proposed the countermeasure based upon modulation index and spectral ratio. The study focused on detecting the far-field recording of the genuine speaker for landline and GSM telephone channel. The ASV spoof 2017 challenge campaign came up with various countermeasures for replay attack detection. The Variable length Energy Separation Algorithm-Instantaneous Frequency Cosine Coefcients (VESA-IFCC) feature set was proposed in [17], to capture the characteristics of natural and replay speech. In same study authors also discussed the effectiveness of VESA-IFCC feature using spectrographic analysis. In [18], authors showed that the importance of high frequency region for the replay spoof detection by considering several frequency ranges for feature extraction. To exploit the characteristics of natural and replay speech, authors in [19] proposed two source-based feature sets, namely, Epoch Features (EF) and Peak to Side lobe Ratio-Mean and Skew (PSRMS). Furthermore, these feature sets are fused at score-level with Instantaneous Frequency Cosine Coefficients (IFCC) [20], Mel Frequency Cepstral Coefficients (MFCC) [21] and Constant Q Cepstral Coefficients (CQCC) [22] to capture the possible complementary information. The major distinguishable factor between natural and replay is that replay speech is passed through several channels as opposed to natural speech. To detect this channel information, Single Frequency Filtering (SFF) approach was proposed in [23]. Authors in [24]–[27] implemented replay spoof detection system using various neural network approaches, such as ensemble learning, ResNet, Bidirectional Short Long Short Term Memory (BLSTM) etc. The best performing system in ASV spoof 2017 challenge was reported in [28], where the authors studied single Convolutional Neural Network (CNN) and combined with Recurrent Neural Network (RNN) approaches.

The Teager Energy Operator (TEO) phase feature set was originally proposed for speaker recognition task [29]. The TEO phase captures the excitation source-related information, which is complementary to speaker-specific information ob-

tained through spectral features, such as CQCC, MFCC, etc. [29]. In addition, TEO phase does not require pre-processing operations, such as framing, windowing, pre-emphasis etc. In TEO phase feature extraction process, the problem of accurate GCI detection was addressed by using singularity detection through wavelet analysis [29]. In this work, we explore the TEO phase feature set for replay spoof detection task. Furthermore, TEO phase feature set fused at score-level with magnitude-based features, namely, CQCC [22], MFCC [21], and Linear Frequency Cepstral Coefficients (LFCC) [30].

II. Teager Energy Operator (TEO) Phase

Various conventional features, such as MFCC, Linear Prediction Cepstral Coefficients (LPCC) assumes that the speech production mechanism is linear in which the airflow propagation through vocal tract is linear plane wave. However, the concomitant vortices are dispersed over entire vocal tract area and the airflow is separated and hence, the assumption of linearity may fail [31], [32]. The actual source of speech production is vortex-flow interactions, these vortex-flow interactions are nonlinear in nature. The TEO is a nonlinear energy tracking operator for signal analysis and to characterize the airflow properties in vocal tract [31]. Considering a fact that energy in producing an acoustical signal (such as speech) is a dependent on its frequency as well as amplitude, Kaiser developed a TEO operator $\psi(n)$ for discrete-time signal $s(n)$ as [33],

$$\psi(n) = \psi\{s(n)\} = s^2(n) - s(n+1)s(n-1). \quad (1)$$

Around Glottal Closure Instants (GCIs), the TEO profile gives higher energy value. Motivated by a study reported in [34], the authors in [29] used phase of an analytic signal obtained from TEO profile of speech frame. The analytic signal $\psi_a(n)$ for TEO profile is given by,

$$\psi_a(n) = \psi(n) + j\hat{\psi}(n), \quad (2)$$

where $\hat{\psi}(n)$ is a Hilbert transform of $\psi(n)$. The Hilbert transform produce the phase shift of 90^o for every frequency component and can be computed as follows,

$$\hat{\psi}(n) = \mathcal{F}^{-1}\left(\hat{\Psi}(\omega)\right), \quad (3)$$

where $\mathcal{F}^{-1}$ is inverse Fourier transform and $\hat{\Psi}(\omega)$ is Fourier transform of $\hat{\psi}(n)$ given as,

$$\hat{\Psi}(\omega) = \begin{cases} -j\Psi(\omega), & if \quad 0 \leq \omega < \pi, \\ j\Psi(\omega), & if \quad -\pi \leq \omega < 0, \end{cases} \quad (4)$$

where $\Psi(\omega)$ denotes Fourier transform of the TEO profile $\psi(n)$. The amplitude envelope of analytic signal also known as Hilbert envelope is given by,

$$a_e(n) = \sqrt{\psi^2(n) + \hat{\psi}^2(n)}. \quad (5)$$

The TEO phase is cosine of the phase of analytical signal $\psi_a(n)$ and computed as,

$$\phi_\psi(n) = \cos\left(\angle\psi_a(n)\right) = \frac{\psi(n)}{a_e}. \quad (6)$$

where $\phi_\psi(n)$ denotes the TEO phase.

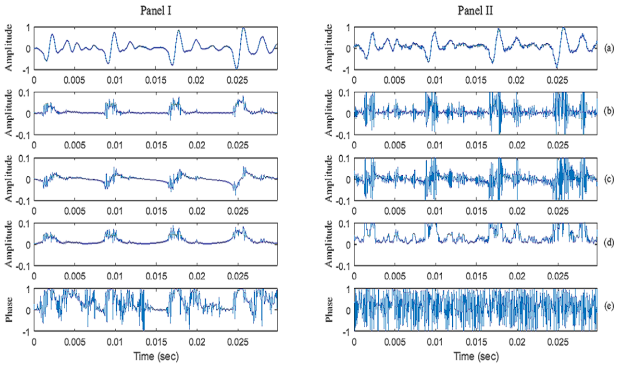

Fig. 1. (a) Voiced speech segment (b) TEO profile (c) Hilbert transform (d) Hilbert envelope (e) TEO phase (Panel I: genuine speech segment, Panel II: corresponding replay speech segment).

The Figure 1 shows the voiced segment of speech signal, its TEO profile, the Hilbert transform of TEO profile, Hilbert envelope and TEO phase for genuine (panel I) and similar analysis for corresponding replay speech (panel II). The Figure 2 shows the similar analysis for speech segment containing silence region followed by voiced region for genuine (panel I) and replay speech (panel II). From Figure 1, it can be observed that the TEO phase plot of the replay speech (panel II) is more fluctuating compared to the genuine speech (panel I) signal in case of voice speech segment. From Figure 2, it can be noticed that the genuine speech (panel I) signal containing silence region gives almost zero TEO phase values for silence region, unlike replay speech (panel II) signal which gives significant TEO phase values in silence region (because small bumps present in Hilbert envelope of silence region).

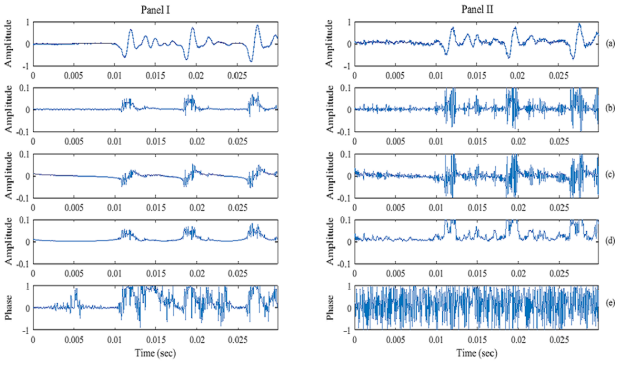

Fig. 2. (a) Speech signal having silence region followed by voiced segment (b) TEO profile (c) Hilbert transform (d) Hilbert envelope (e) TEO phase (Panel I: genuine speech segment, Panel II: corresponding replay speech segment).

The another observation is that although TEO profile indicates energy, it can have negative values (as can be observed

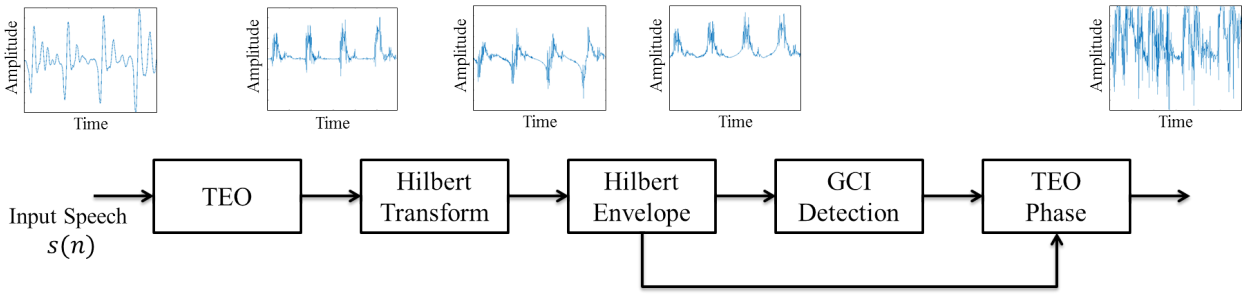

Fig. 3. Functional block diagram to extract TEO phase feature set. After [29].

from Eq. (1)) and have higher energy values when vocal tract gets sudden impulse-like excitation. From Figure 1 and Figure 2, it can be observed that for genuine speech TEO profile gives higher values near GCIs, however, for replay speech TEO profile gives higher values around GCIs as well as other locations. This may be due to the noise present in replay speech signal which contribute to running estimate of energy. It is also observed that the TEO phase has better correlation with input speech signal.

From Figure 2, it is clear that for silence region of genuine speech TEO profile has approximately zero energy and hence Hilbert envelope and TEO phase also have zero energies. However, in the replay speech presence of some noisy samples results in spurious TEO values and hence Hilbert envelope and TEO phase have non-zero energies. We also observed that the energy values at GCIs for replay speech gets amplify compared to genuine speech signal, this may be due to fact that replay speech signal is noisy genuine speech signal (replay can be modeled as convolution of genuine speech signal with impulse response of intermediate devices, impulse response of recording and playback environment). From these observation, we can see the potential of the TEO phase information for replay spoof detection.

Figure 3 shows the schematic diagram to estimate the TEO phase feature set. Here, first the TEO profile of the input speech signal is computed using Eq. (1). The Hilbert envelope of the TEO profile is computed from analytic signal of TEO profile using Eq. (5). The feature vector is formed by taking B blocks each of N_d samples of TEO Phase with some shift at the GCI, however this requires exact location of GCIs. Figure 1 and Figure 2 shows that the TEO profile is blunted and hence, the better singularity detection algorithm (for GCIs estimation) is required. The multiscale edge detection can be done using Canny edge detector which is equivalent to wavelet modulus maxima using Gaussian kernel. For singularity detection, wavelet analysis is used, to do this first local fluctuations in Hilbert envelope needs to removed. To get rid of these fluctuations local mean smoothing followed by its wavelet transform of Hilbert envelope is done. The wavelet transform of a signal can be expressed as multiscale differential operator [35]. In [36], it is reported that all the singularities present in signal can be detected using wavelet transform modulus maxima at finer scales. This property of signal is used for GCI detection in TEO phase feature extraction. The derivatives of the Gaussians are widely used in numerical computations to make sure all the maxima line propagate up to the fine scales (pp.177-178, [35]). As TEO profile is calculated for entire input speech signal, it avoids voiced/unvoiced detection, pre-emphasis, framing and have less computation cost.

III. Experimental Setup

A. Database and Classifier

All the experiments are performed on the ASV spoof 2017 challenge database. All speech utterances have a resolution of 16 bits per sample and sampling frequency of 16 kHz. The details of the database can be found in [13]. All the systems are implemented with GMM classifier with appropriate Gaussian components. Two GMMs are trained for genuine and spoof class using only training set of ASV spoof 2017 challenge database.

B. Feature Extraction

System S_1 built with TEO phase feature set. The 6 blocks each of 40 samples with one sample shift of TEO Phase at the GCI is taken to form *40*-dimensional (D) feature vector. The GCIs are estimated using Hilbert envelope and *1*-D Canny operator. The system S_2 is built with *90*-D CQCC features that comprise of the zeroth coefficient, 29-static, 30-Δ, and 30-$\Delta\Delta$ coefficients. The minimum frequency set to 15 Hz and maximum frequency to 8 kHz, the number of bins per octave set to 96.

TABLE I
Summary of the experimental setups of the systems *(FD: Feature Dimension)*

System	Feature Set	FD	No. of Gaussians
S_1	TEO phase	40	256
S_2	CQCC *(Baseline)*	90	512
S_3	MFCC	39	512
S_4	LFCC	180	512

System S_3 developed with *39*-D (13-static + 13-Δ + 13-$\Delta\Delta$) MFCC features. Total 40 triangular filters along with the Hamming window of 20 ms duration and 10 ms shift are used for the feature extraction process. System S_4 is based on LFCC. The LFCC features are extracted with 60

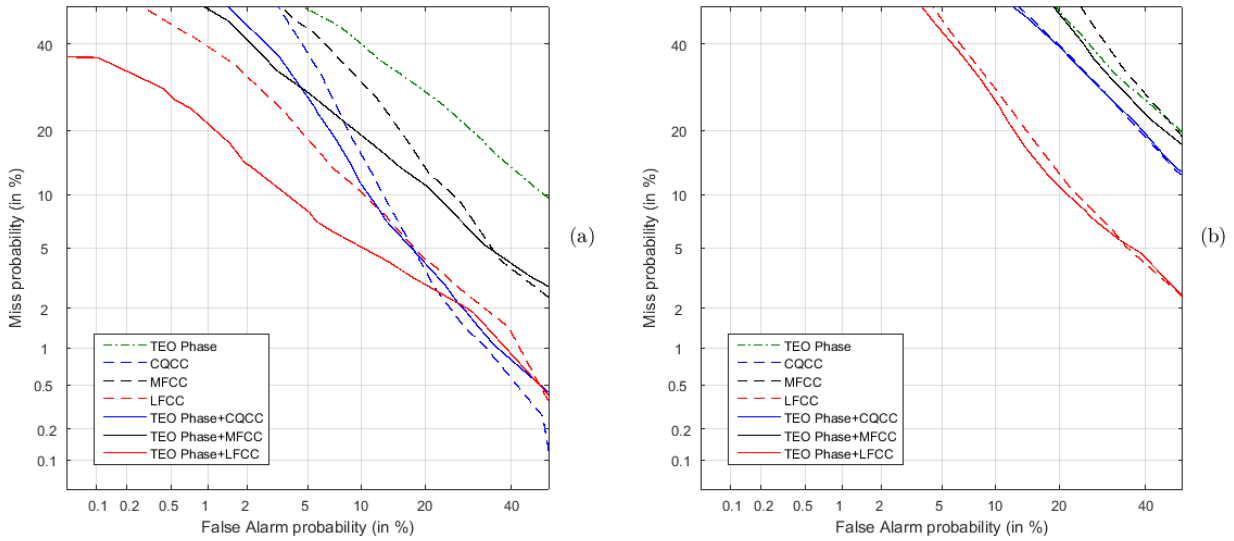

Fig. 4. DET curves for (a) development set and (b) evaluation set.

triangular filters and using frame length of 20 ms with 50 % overlap. Extracted features are appended with 60-Δ and 60-$\Delta\Delta$ coefficients resulting *180*-D feature vector. Table I summarizes the experimental setup used for development of spoof detection system.

IV. Experimental Results

The experimental results of the replay spoof detection on development and evaluation set are given in Table II. From results, it can be observed that the individual TEO phase feature do not perform well, however, when they are fused with magnitude-based features, the system performance improves substantially. This indicates that the TEO phase feature contain complementary information to the magnitude-based features.

TABLE II
Results for development and evaluation set

System	EER (%)	
	Development	**Evaluation**
TEO phase	23.91	31.34
CQCC	11.89	28.92
MFCC	17.27	34.02
LFCC	10.28	16.80
TEO phase+CQCC	10.62	28.74
TEO phase+MFCC	14.56	31.28
TEO phase+LFCC	**6.57**	**15.39**

'+' indicates score-level fusion

The organizers of the ASV spoof challenge provided CQCC-GMM as baseline system with an EER of 28.92 % on evaluation set of database. The standalone spoof detection system built with TEO phase, MFCC and LFCC gives a result of 31.34 %, 34.02 %, and 16.80 %, respectively, on evaluation set. When TEO phase feature set fused with CQCC, MFCC, and LFCC EER gets reduced by 0.18 %, 2.74 %, and 1.41 %, respectively, compared to the corresponding magnitude-based feature sets. This improvement in system performance points out that the presence of TEO phase along with magnitude information strengthens the spoof detection system. Figure 4 shows the DET curves for development and evaluation set of ASV Spoof 2017 Challenge database.

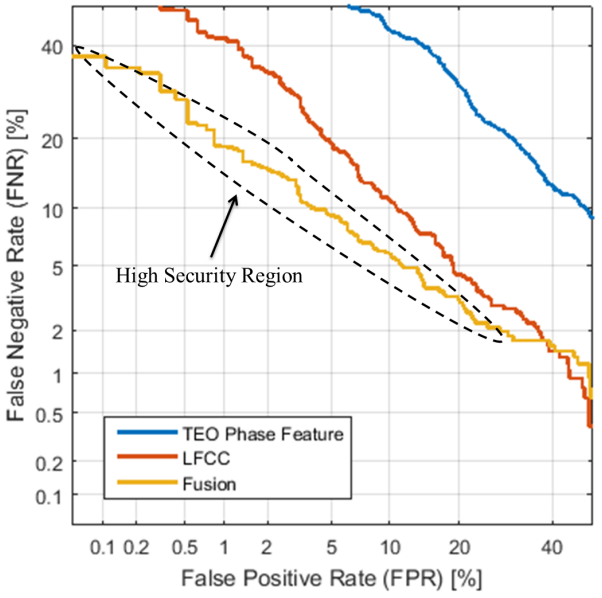

Fig. 5. DET curves for TEO phase, LFCC and their fusion for development set *(highlighted portion indicated deviation towards high security region)*.

Figure 5 shows the DET curve for TEO phase, LFCC and their score-level fusion for development set, similar curves

were observed for CQCC, MFCC and IMFCC. From the DET curves, it is observed that when magnitude-based features fused with TEO phase feature the DET curve deviates towards vertical-axis more compared to the horizontal-axis i.e. the probability of false acceptance is less, however probability of false rejection is comparatively high. This indicates that the TEO phase feature capture the information required for designing high security replay spoof detection system for ASV. TEO phase feature set detects the spoofed speech very efficiently and does not allow the attacker to access the ASV system easily, which is very important in practical applications.

V. Summary and Conclusions

This paper explore the significance of TEO phase feature set for replay spoof detection task. We observed that the TEO phase plots seems to be very noisy for replay speech compared to natural speech. In this work, we have investigated TEO phase feature performance with CQCC, MFCC and LFCC feature. We observed that the TEO phase feature gives the complementary information to the speaker-specific information provided by CQCC, MFCC and LFCC feature sets. We also observed that the TEO phase feature provide a information which deviates DET curve towards high security reason than high user convenience region, indicating that TEO phase efficiently detects replayed speech. In future, Variable length Teager Energy Operator (VTEO) phase can be used with magnitude information for better system performance. Neural network based classifiers like CNN can also be used to enhance the system performance.

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] 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, China, 2004, pp. 145–148.

[3] 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*, Melbourne, Australia, 2005, pp. 15–21.

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

[5] N. Evans, F. Alegre, Z. Wu, and T. Kinnunen, "Anti-spoofing, voice conversion," *Encyclopedia of Biometrics*, pp. 115–122, 2015.

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

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[8] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 1211–1214.

[9] 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.

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

[11] H. A. Patil and K. K. Parhi, "Variable length teager energy based mel cepstral features for identification of twins," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2009, pp. 525–530.

[12] 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-ASC)*, 2014, pp. 1–5.

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

[14] 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.

[15] 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.

[16] J. Villalba and Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE International Carnahan Conference on Security Technology (ICCST)*, Barcelona, Spain, 2011, pp. 1–8.

[17] 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.

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

[19] 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.

[20] 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.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in Speech Rrecognition*. Elsevier, 1990, pp. 65–74.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] H. A. Patil and K. K. Parhi, "Development of TEO phase for speaker recognition," in *IEEE, International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2010, pp. 1–5.

[30] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear *vs.* mel frequency cepstral coefficients for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, HI, USA, 2011, pp. 559–564.

[31] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*. Springer, 1990, pp. 241–261.

[32] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.
[33] 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, NM, USA, 1990, pp. 381–384.
[34] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.
[35] S. Mallat, *A Wavelet Tour of Signal Processing*. Second Edition, Academic press, 1999.
[36] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 617–643, 1992.