Interspeech 2018
2-6 September 2018, Hyderabad

Effectiveness of Speech Demodulation-Based Features for Replay Detection

Madhu R. Kamble, Hemlata Tak and Hemant A. Patil

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

{madhu_kamble, hemlata_tak, hemant_patil}@daiict.ac.in

Abstract

Replay attack presents a great threat to Automatic Speaker Verification (ASV) system. The speech can be modeled as amplitude and frequency modulated (AM-FM) signals. In this paper, we explore speech demodulation-based features using Hilbert transform (HT) and Teager Energy Operator (TEO) for replay detection. In particular, we propose features, namely, HT-based Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) Cosine Coefficients (i.e., HT-IACC and HT-IFCC) and Energy Separation Algorithm (ESA)-based features (i.e., ESA-IACC and ESA-IFCC). For adapting instantaneous energy w.r.t given sampling frequency, ESA requires 3 samples whereas HT requires relatively large number of samples and thus, ESA gives high time resolution.The experiments were performed on ASV spoof 2017 Challenge database for replay spoof speech detection (SSD).The experimental results shows that ESA-based features gave lower EER. In addition, linearly-spaced Gabor filterbank gave lower EER than Butterworth filterbank. To explore possible complementary information using amplitude and frequency, we have used score-level fusion of IA and IF. With HT-based feature set, the score-level fusion gave EER of 5.24 % (dev) and 10.03 % (eval), whereas ESA-based feature set reduced the EER to 2.01 % (dev) and 9.64 % (eval).
Index Terms: Spoofing, Hilbert transform, Teager energy operator, energy separation algorithm.

1. Introduction

A voice biometrics system deals with the recognition of a speaker through his/her voice and also with their humming sounds [1–3]. However, the recent advances in speech technologies have posed a great threat to the Automatic Speaker Verification (ASV) system with various *spoofing* attacks [4]. There are five types of spoofing attacks, namely, impersonation, replay, speech synthesis (SS), voice conversion (VC) and twins [5], [6]. The ASV spoof 2017 Challenge was focused on replay spoof attack. The organizers provided Constant Q Cepstral Coefficients (CQCC) as a baseline feature set with a simple Gaussian Mixture Model (GMM) as classifier [7]. Various countermeasures were proposed for detecting the replay spoofed speech [8]. Few of the countermeasures focused on the high frequency spectral information, feature normalization and the representation learning [9–12]. Recently, we have proposed Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (ESA-IFCC) [13] and its variable length version (VESA-IFCC) feature set for SSD task for SS, VC, and replay detection [14]. The information of Instantaneous Amplitude (IA) for each subband signal was ignored and only Instantaneous Frequency (IF) was considered using Teager Energy Operator (TEO) [13–15]. In this paper, we propose to exploit IA information in addition to IF with linearly-spaced Gabor filterbank. Furthermore, ESA-based demodulation approach is compared with the other existing demodulation technique using Hilbert transform (HT) [16]. In particular, IA and IF-based features extracted from HT-based Energy Separation Algorithm (HT-ESA) and Teager Energy Operator-based ESA (TEO-ESA) are used. To adapt the instantaneous energy w.r.t given sampling frequency, HT requires a large (by a factor *40-50* number of samples and hence, it has relatively poor time resolution [17]. While ESA requires only three samples (for a given sampling frequency) to adapt the instantaneous energy and thus, has excellent time resolution.

In this paper, results were compared for both the Gabor and Butterworth filterbank and found the importance of filterbank parameters, such as (shape of filter, numbers of subband filters, choice of bandwidth) for spoof speech detection (SSD) task. Furthermore, to exploit high frequency information and to overcome channel mismatch conditions, pre and post-processing (i.e., pre-emphasis filter and cepstral mean normalization) of a speech signal is used to compute the features. The proposed feature sets were compared with the Constant Q Cepstral Coefficients (CQCC) (baseline system provided by the challenge organizers), Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC) on the ASV spoof 2017 Challenge database.

2. Demodulation using HT and TEO

2.1. Hilbert Transform (HT)

One of the demodulation method was proposed using HT [16]. It is implemented by doubling the spectrum for positive frequencies and setting it zero for the negative frequencies (i.e., via analytic signal generation) [16], [18]. Therefore, the energy of the signal does not change (i.e., Parseval's energy conservation). The HT estimates frequency and amplitude envelope function of a monocomponent signal. The IA, $a(t)$, instantaneous phase, $\phi(t) = tan^{-1}\frac{x_i(t)}{x_r(t)}$, and IF, $\phi'(t)$ obtained from HT-ESA are given as:

$$a(t) = \sqrt{x_r^2(t) + x_i^2(t)}, \tag{1}$$

where $x_r(t)$ and $x_i(t)$ are the real and imaginary parts of the analytic signal. The instantaneous or analytic phase and frequency is given by:

$$\phi(t) = tan^{-1}\frac{x_i(t)}{x_r(t)}. \tag{2}$$

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

Though HT tracks frequency and amplitude variations, however, it is based on Fourier transform (FT) and hence, have the limitations of stationarity and linearity of Fourier analysis [18], [19]. In addition, HT-based IF estimation requires computationally complex task of phase unwrapping in Eq. (3) [20].

10.21437/Interspeech.2018-1675

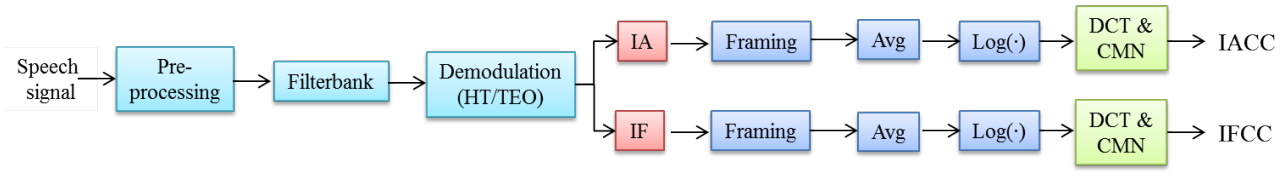

Figure 1: *Schematic block diagram of proposed demodulation-based feature sets.*

Moreover, unwrapping may not be a unique phenomenon [21]. To alleviate this, recent approach was proposed in [20]. However, this approach also requires block-based Discrete Fourier Transform (DFT) processing of speech that requires more than 3 samples. This motivated us to exploit high resolution TEO, for tracking IA and IF [22], [23].

2.2. Teager Energy Operator (TEO)

An algorithm derived by Teager uses a nonlinear energy tracking operator for discrete-time speech signal [24, 25]. The TEO, $\Psi\{\cdot\}$ for discrete-time signal, $x[n] = A\cos(\Omega n + \phi)$, is given as:

$$\Psi\{x(n)\} = x^2(n) - x(n-1)x(n+1) \approx A^2\Omega^2. \quad (4)$$

Speech signal can be modeled as Amplitude and Frequency Modulation (AM-FM) signal, since speech signal have variable amplitude with variations in frequency. Consider an AM-FM discrete-time form of signal, $x[n] = a[n]\cos(\phi[n])$ [22]:

$$x[n] = a[n]\cos\left[\Omega_c n + \Omega_m \int_0^n q[k]dk + \theta\right], \quad (5)$$

with variable IF, $\Omega[n] = \frac{d}{dn}\phi[n] = \Omega_c + \Omega_m q[n]$. The TEO can approximately estimate the squared product of the amplitude, $a[n]$, and IF, $\Omega[n]$ signals, i.e.,

$$\Psi\left(a[n]\cos\left(\int_0^n \Omega[m]dm + \theta\right)\right) \approx a^2[n]\Omega^2[n]. \quad (6)$$

To estimate the individual contribution of amplitude, $a[n]$, and frequency, $\Omega[n]$ Maragos *et. al.* [22], [23] developed an ESA using nonlinear energy operator. The energy of a speech signal is a function of both $a[n]$ and $\Omega[n]$ [26]. The IA, $a[n]$, and IF, $\Omega[n]$, at any time instant of modulated signal is given as [18]:

$$a[n] \approx \frac{2\Psi_d\{x[n]\}}{\sqrt{\Psi_d\{x[n+1]) - x[n-1]\}}}, \quad (7)$$

$$\Omega[n] \approx arcsin\sqrt{\frac{\Psi_d\{x[n+1] - x[n-1]\}}{4\Psi_d\{x[n]\}}}. \quad (8)$$

The key advantages of this ESA algorithm is that it doesn't require complex task of phase unwrapping (as required in HT-based approach Eq. (3)) and in addition only three samples are required to get $a[n]$ and $\Omega[n]$ and thus avoiding the need of block-based processing of speech [20].

3. Proposed Feature Set

Figure 1 shows the block diagram of our proposed speech demodulation-based feature set using HT and TEO. Here, the input speech signal is passed through a pre-emphasis filter which balance the lower and higher frequencies of a speech signal (i.e., flattening the magnitude spectrum) [27]. Its system function is $H(z) = 1 - az^{-1}$. where 'a' (filter coefficient) is a constant with a typical value of *0.97* [27]. The higher formants, such as F_3 and F_4 are used for speaker discrimination that are present in the higher frequency regions [28]. These higher frequency spectral regions are recently found to be important for replay SSD task [29]. The TEO works on single component signal and the speech signal is a multicomponent AM-FM signal and thus, bandpass filtering is needed to isolate each component of a speech signal. Thus, Gabor filterbank is used as a multi-band filtering to separate the signal component in the temporal-domain. The pre-processed speech signal is passed through a Gabor filterbank to obtain (*N=40*) subband filtered signals. 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 of Gaussian is a Gaussian) [30]. The impulse response, $g(t)$, and frequency response, $G(\omega)$, of a Gabor filter is given as [23], [26] :

$$g(t) = exp(-b^2t^2)\cos(\omega_c t), \quad (9)$$

$$G(\omega) = \frac{\sqrt{\pi}}{2b}\left[exp\left(-\frac{(\omega-\omega_c)^2}{4b^2}\right) + exp\left(-\frac{(\omega+\omega_c)^2}{4b^2}\right)\right], \quad (10)$$

where ω_c is the center frequency (in Hz) of the filter and b is a parameter for controlling the bandwidth of a filter. The Gaussian shape of $G(\omega)$ avoids producing side lobes that could produce the false pulses in the output of demodulation (HT/ TEO). This narrowband filtered signal is then passed through both demodulation techniques, i.e., HT-ESA and TEO-ESA. The estimated IA and IF profiles are passed through frame blocking and averaged over a short window of *20* ms and with a shift of *10* ms. The Discrete Cosine Transform (DCT) is used to obtain a low-dimensional feature representation. These feature sets were again post-processed with Cepstral Mean Normalization (CMN) technique to overcome the channel mismatch/ distortion between the training and testing conditions [31–33]. Furthermore, the 40 static coefficients of IA and IF were retained and appended along with their Δ and $\Delta\Delta$ features to get higher-dimensional feature set.

3.1. Filterbank Used

A single speech signal is modeled as the exponentially damped AM-FM signal and the speech signal is a sum of the AM-FM signal [34]. Before applying the demodulation technique, we need to extract the resonance characteristics of the signal through bandpass filtering. In the earlier studies, it was observed that with bandpass filtering of the original signal, the (single or double) spike at the jump instant are converted to smooth sinusoidal curves [17]. The bandpass filtering actually filters out the higher frequency components of the amplitude envelope and instantaneous frequency. We have extracted our feature set using two filterbanks, namely, Butterworth and Gabor. In our earlier work, we have used Butterworth filterbank with linearly-spaced filters across entire frequency range [13]. In this paper, we have compared results with both linearly-spaced Butterworth and Gabor filterbanks. The use of linear scale makes the feature extraction process more reliable as all

the filters have almost equal bandwidths [15]. The effect of the Gabor filterbank is to smooth out the spikes and the abrupt jumps (if any) of the original estimates where high frequency components are preserved [17]. The spectral energy density obtained with 40 number of subband filtered signals is shown in Figure 2. The time-domain speech signal of the utterance, *"Actions speak louder than words,"* is shown in Figure 2 (a), whereas Figure 2 (b) and Figure 2 (c) shows the spectral energy density obtained from Butterworth and Gabor filterbanks (both having 40 subband filters), respectively. The spectral energies obtained from the Gabor filterbank preserves more of the lower frequency information (highlighted with the dotted box) than the spectral energies obtained from the Butterworth filterbank. These lower frequency information provides the lower formant information (i.e., F_1 and F_2) that contains the information about the message present in the signal. The higher formants (i.e.,F_3 and F_4) present in higher frequency region is also preserved because of linear frequency scale. These higher formants are important for speaker discrimination [35]. The higher formants are preserved because of linearly-spaced Gabor filterbank.

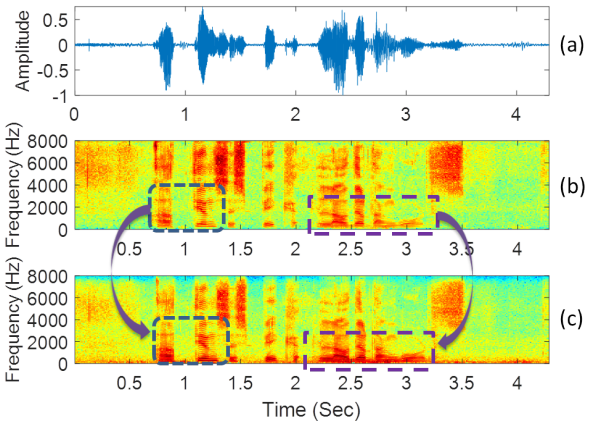

Figure 2: *Comparison of spectral energy density between Butterworth and Gabor filterbank. (a) time-domain speech signal, spectral energy density of (b) Butterworth filterbank, and (c) Gabor filterbank obtained with 40 subband filtered signals.*

4. Experimental Setup

The experiments were performed on ASV spoof 2017 Challenge database which was focused only on replay spoof attacks [36]. The details of statistics of database, recording conditions, playback, and recording devices are provided in [8]. The proposed features, i.e., HT-IACC, HT-IFCC, ESA-IACC, and ESA-IFCC were extracted from *40* narrowband filtered signals with *40* static coefficients appended with Δ and $\Delta\Delta$ resulting in the *120*-dimensional feature vector. The Gabor filterbank was computed with linear frequency scale from F_{min}= *10* Hz and F_{max}= *8000* Hz. The feature dimension for MFCC was kept as 39-D (static+Δ+$\Delta\Delta$) and for CQCC, it was 90-D (static+Δ+$\Delta\Delta$). We have used Gaussian Mixture Model (GMM) as classifier with 512 number of mixtures. The decision of the speech signal of being genuine or replayed is based on the scores of Log-Likelihood Ratio (LLR) [3]:

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \quad (11)$$

where $P(X|H_0)$ and $P(X|H_1)$ are the likelihood scores for genuine and replay trials (with hypothesis H_0 and H_1), respectively. To explore the complementary information of proposed feature set, score-level fusion was performed with CQCC, MFCC and LFCC feature set as per Eq. (12):

$$LLK_{fused} = \alpha LLK_{feature1} + (1-\alpha)LLK_{feature2}, \quad (12)$$

where $LLK_{feature1}$ is a log-likelihood score of CQCC, MFCC and LFCC $LLK_{feature2}$ is the score of our proposed feature set. The fusion parameter (α) lies between $0 < \alpha < 1$ to decide the weight of scores. The performance of a voice biometric system is generally calculated by the Equal Error Rate (EER). It corresponds to a threshold at which the False Acceptance (FA) rate is equal to the False Rejection (FR) rate. The FA and FR of a verification system define different operating points on the Decision-Error Trade-off (DET) curve [37].

5. Experimental Results

Table 1 shows the result of proposed feature set extracted from the Gabor filterbank along with the effect of logarithm method. The AM feature set extracted from HT and ESA with log performed better than those extracted from without log. On other hand, FM feature set when extracted with log did not reduced the EER, as the IF have more fluctuations because of which the dynamic range varies more as compared to the IA feature and thus, the FM feature set with log did not gave better results.

Table 1: *Effect of log on proposed feature set with Gabor filterbanks on dev and eval set*

Feature Set	Without Log		Log	
	Dev	Eval	Dev	Eval
HT-IACC	32.76	39.90	**07.27**	**12.12**
HT-IFCC	**14.07**	**14.62**	31.81	38.65
ESA-IACC	29.05	33.94	**06.48**	**12.00**
ESA-IFCC	**04.12**	**12.79**	37.35	28.29

5.1. Results on Different Filterbank Used

The results of the AM-FM features with Butterworth and Gabor filterbanks are shown in Table 2. The overall performance with Gabor filterbank shows better results than Butterworth filterbank. The ESA-IFCC feature set gave an EER of 4.12 % on dev set with Gabor filterbank while on eval, it is 12.79 %, which is much lower EER than the features extracted from the Butterworth filterbank.This indicates that the choice of a linearly-spaced Gabor filterbank indeed helps for feature extraction.

Table 2: *Results of proposed feature set with Butterworth and Gabor filterbanks on dev and eval set*

Feature Set	Butterworth		Gabor	
	Dev	Eval	Dev	Eval
HT-IACC (A)	**09.74**	**19.27**	07.27	12.12
HT-IFCC (B)	15.41	39.40	14.07	14.62
ESA-IACC (C)	17.59	21.43	06.48	**12.00**
ESA-IFCC (D)	18.82	28.69	**04.12**	12.79

5.2. Results with Score-Level Fusion

The database organizers provided a baseline system with CQCC as the feature set [8]. To explore the possible complementary information captured by various feature sets, we have used their score-level fusion. We have compared our results with CQCC, MFCC and LFCC feature sets, LFCC is used for comparison of our results because the proposed feature extraction was done with linear scale. The results of score-level fusion with our proposed feature sets obtained from Gabor filterbank is shown in Table 3. The results on dev set reduced the EER in almost every

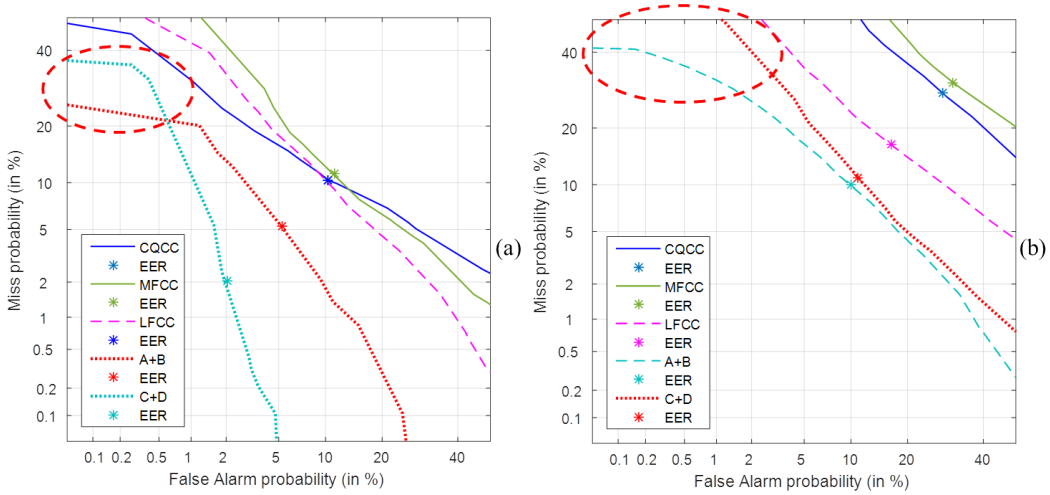

Figure 3: *Individual DET curves of CQCC, MFCC, LFCC, A+B (score-level fusion of HT-IACC+HT-IFCC) and C+D (score-level fusion of ESA-IACC+ESA-IFCC) feature sets on (a) dev set and (b) eval set. Dotted circle indicated better performance for SSD task*

case whereas on eval set, except for the few cases, it reduced the individual EER. On dev set, score-level fusion EER of CQCC, MFCC and LFCC feature set with ESA-IFCC reduced (from 4.12 %) to 2.35 %, 2.41 % and 2.82 %, respectively. While on eval set, the fusion reduced the EER almost to 12.02 % (CQCC), 12.79 % (MFCC) and 11.16 % (LFCC). On the other hand, for ESA-IACC feature set, the EER via score-level fusion was reduced for only few cases on both dev and eval set. Better results were obtained via fusion with ESA-based technique than its HT-based counterpart.

Table 3: *Results with score-level fusion of CQCC, MFCC, LFCC and proposed feature sets (A-D)*

Feature Set	CQCC		MFCC		LFCC	
	Dev	Eval	Dev	Eval	Dev	Eval
A	03.74	11.47	04.39	12.12	07.19	12.12
B	05.33	14.16	05.80	14.62	07.60	11.62
C	04.07	**11.30**	04.12	**12.00**	06.48	12.00
D	**02.35**	12.02	**02.41**	12.79	**02.82**	**11.16**

A: HT-IACC, B: HT-IFCC, C: ESA-IACC, D: ESA-IFCC

To explore the possible complementary information present in the proposed HT and ESA-based feature sets, we have fused the individual IA and IF-based information. The proposed features are amplitude and frequency-based, (i.e., the amplitude-based feature set do not have any information of frequency and vice-versa). This fusion indeed helps to reduce the EER further for both HT and ESA-based feature sets. The results of fusion of these proposed feature sets is shown in Table 4. When the HT-IACC and HT-IFCC feature sets were fused at a score-level, on dev set the EER reduced to 5.24 % and on eval set, it reduced to 10.03 %. Similarly, when ESA-IACC and ESA-IFCC feature sets were fused at score-level, the EER further decreased to 2.01 % on dev set and on eval set, it gave the better result of 9.64 %. The overall results are summarized in Table 4. The baseline system using CQCC provided by the organizers of ASV Spoof 2017 Challenge has much higher EER (28.48 %) on eval set. The results with MFCC and LFCC on eval set gave an EER of 31.31 % and 16.62 %, respectively. Our proposed best result obtained after the score-level fusion of ESA-IACC and ESA-IFCC feature set with an EER of 2.01 % on dev set and 9.64 % on eval set. The performance is also shown by the DET curves in Figure 3 (a) for dev set and Figure3 (b) for eval set with different feature sets CQCC, MFCC, LFCC, score-level fusion (HT-IACC+HT-IFCC) and score-level fusion (ESA-IACC+ESA-IFCC). On dev and eval set, score-level fusion (ESA-IACC+ESA-IFCC) shows relatively better performance for all the operating points of DET curve and have significantly lower false alarm and miss probabilities in the DET curve when compared to CQCC, MFCC and LFCC feature sets (shown by dotted circle).

Table 4: *Comparison of best proposed feature set with other feature set on dev and eval set*

Feature Set	Dev	Eval
CQCC (Baseline)	10.35	28.48
MFCC	11.21	31.30
LFCC	10.58	16.62
A+B	05.24	10.03
Proposed best result (C+D)	**02.01**	**09.64**

A: HT-IACC, B: HT-IFCC, C: ESA-IACC, D: ESA-IFCC

6. Summary and Conclusions

In this study, we studied the demodulation-based features to detect natural *vs.* replayed spoofed speech. The computation of IA and IF from HT-ESA and TEO-ESA was affected by the parameters of filter, namely, shape of filter, choice of bandwidth, time resolution, etc. In particular, linearly-spaced Gabor filterbank performed better than its Butterworth counterpart. The proposed ESA-based feature set gave lower EER than the HT-based feature set. To explore the complementary information of proposed feature set, we fused them with existing feature sets, namely, CQCC, MFCC and LFCC. The results obtained after score-level fusion gave relatively lower EER than the individual EER. Furthermore, when the proposed feature sets itself when fused at a score-level (i.e., ESA-based IACC+IFCC) gave the best lower EER. Our future work includes the study of reverberation, frequency response characteristics of the replayed device in the higher frequency regions and its relevance for SSD task.

7. Acknowledgments

The authors would like to thank University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF) and authorities of DA-IICT Gandhinagar.

8. References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] H. A. Patil and M. C. Madhavi, "Combining evidences from magnitude and phase information using VTEO for person recognition using humming," *Computer Speech & Language, In Press*, Sept. 2017.

[3] D. A. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," in *The MIT Lincoln Laboratory Journal*, 1995, pp. 173–192.

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

[5] 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.

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

[7] 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.

[8] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2–6.

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

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

[11] 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 2017*, Stockholm, Sweden, 2017, pp. 17–21.

[12] 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.

[13] M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," in *European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 116–120.

[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] M. R. Kamble and H. A. Patil, "Effectiveness of mel scale-based ESA-IFCC features for classification of natural vs. spoofed speech," in *B.U. Shankar et. al. (Eds.) PReMI, Lecture Notes in Computer Sciance (LNCS)*. Springer, 2017, pp. 308–316.

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

[17] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, no. 1, pp. 95–120, 1994.

[18] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 1^{st} edition, Pearson Education India, 2015.

[19] 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.

[20] 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.

[21] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 2, pp. 170–177, 1977.

[22] P. Maragos, J.F. Kaiser and T.H. Quatieri , "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.

[23] 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.

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

[25] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *W. J. Hardcastle and A. Marchal (Eds.), Speech Production and Speech Modeling, Kluwer Academic Publishers*, vol. 55, pp. 241–261, 1990.

[26] 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.

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

[28] Q. Lin, E.-E. Jan, C. Che, D.-S. Yuk, and J. Flanagan, "Selective use of the speech spectrum and a vqgmm method for speaker identification," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4. IEEE, 1996, pp. 2415–2418.

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

[30] S. Mallat, *A Wavelet Tour of Signal Processing*. 3^{rd} Edition, Academic press, 1999.

[31] 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.

[32] M. Alam, P. Ouellet, P. Kenny, and D. OShaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," *Advances in Nonlinear Speech Processing*, pp. 246–253, 2011.

[33] 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.

[34] 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*, 1991, pp. 421–424.

[35] 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)*, Hawaii, USA, 2011, pp. 559–564.

[36] K.-A. Lee, A. Larcher, G. Wang *et al.*, "The reddots data collection for speaker recognition." in *INTERSPEECH*, 2015, pp. 2996–3000.

[37] A. Martin, "The DET curve in assessment of decision task performance," in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 1895–1898.