Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Analysis of Teager Energy Profiles for Spoof Speech Detection

Madhu R. Kamble[1], *Pulikonda Aditya Krishna Sai*[2], *Maddala V. Siva Krishna*[3], *Hemant A. Patil*[1]

[1] Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India. [2] Arizona State University, USA. [3] Mettl, India.

{madhu_kamble, hemant_patil}@daiict.ac.in, {201551013, 201551045}@iiitvadodara.ac.in

Abstract

The recent advances in the technologies pose a threat to the Automatic Speaker Verification (ASV) systems using different spoofing attacks, such as voice conversion (VC), speech synthesis (SS), and replay. To enhance the security of the ASV system, the need raised for the development of efficient anti-spoofing algorithms to detect spoof speech signals from natural signal. In this paper, we exploit Teager energy-based features for spoof speech detection (SSD) task. The Teager energy profiles computed for natural, VC, SS, and replay signals show the changes around the Glottal Closure Instants (GCIs). In particular, for SS signal, the bumps are very smooth compared to the natural signal. These variations around GCI of Teager energy profiles helps to discriminate the spoof signal from natural counterparts. The experiments are performed on ASVspoof 2015 and BTAS 2016 challenge databases. The Teager energy-based feature set, i.e., Teager Energy Cepstral Coefficients (TECC) performs well for S1-S9 spoofing algorithms obtaining average EER of 0.161 % (however, not for S10, where EER is 58.14 %) whereas state-of-the-art features, namely, Cochlear Filter Cepstral Coefficients-Instantaneous Frequency (CFCC-IF), and Constant-Q Cepstral Coefficients (CQCC) gave an EER of 0.39 % and 0.163 %, respectively. It is interesting to note that significant negative result by proposed feature set to S10 *vs.* natural speech confirms capability of TECC to represent characteristics of airflow pattern during natural speech production. Furthermore, the experiments performed on BTAS 2016 challenge dataset, gave 2.25 % EER on development set. On evaluation set, TECC feature set gave Half Total Error Rate (HTER) of 3.7 % which is the metric provided by the challenge organizers and thus, overcoming the baseline by a noticeable difference of 3.16 %.

keywords: Spoof, Replay, Teager Energy Operator, Teager Energy Profiles.

1. Introduction

Automatic Speaker Verification (ASV) or voice biometrics system gives the access to the authentic user by using the voice of the claimed speaker [1]. It reduces the risk that relates to the authentication which requires the passwords or sharing of sensitive data. However, ASV systems are vulnerable to various kinds of spoofing attacks, namely, speech synthesis (SS) [2], voice conversion (VC) [3], replay [4], impersonation [5], and twins [6]. The ASV systems can be secured by developing independent Spoof Speech Detection (SSD) system [7, 8, 9]. Due to the recent advances in technology, synthetic, and converted voices have excellent quality (including naturalness), and resembles close to human speech. These machine-generated speech samples generally use the techniques that concentrate on mapping the spectral characteristics. These synthetic, and voice converted speeches can be deliberately used to deceive the ASV systems [10].

The task of the first ASVspoof 2015 spoofing and countermeasures challenge was to design an independent SSD countermeasure that discriminates the spoof speech from the natural speech [7]. In ASVspoof 2015 challenge, various countermeasures were proposed, such as Constant-Q Cepstral Coefficients (CQCC) [11], Linear Frequency Cepstral Coefficients (LFCC) [12], and Cochlear Filter Cepstral Coefficients-Instantaneous Frequency (CFCC-IF) [13]. Other countermeasures include relative phase shift and short-time Fourier transform phase-based features [14, 15, 16, 17]. Neural network-based approaches were also used, such as Convolutional Neural Networks (CNN), and Recurrent Neural Network (RNN) along with front-end features, namely, Teager Energy Operator (TEO) Critical Band Autocorrelation Envelope (TEO-CB-Auto-Env), Perceptual Minimum Variance Distortion less Response (PMVDR), and raw spectrograms [18]. The frame-level and sequence-level features were extracted using Deep Neural Network (DNN) and RNN in [19]. Bottleneck features extracted from the DNN hidden layers were also used with GMM classifier in [20]. We proposed to use Convolutional Restricted Boltzmann Machine (ConvRBM) for auditory filterbank learning that performed better than the traditionally handcrafted filterbank structure used for SSD task.

Compared to the first anti-spoofing challenge, i.e., ASVspoof 2015 [7], which focuses on the synthetic and converted speech attacks (termed as '*Logical Access (LA)*' attacks), the new dataset was released (BTAS 2016), and the competition hubs on *replay attacks* (the first dataset focusing on replay attacks) [22, 23, 24]. The BTAS 2016 competition used the publicly available AVspoof database [25] which provides various presentation attacks that are commonly referred to as *replay attacks*. They used SS and VC as presentation attack wherein natural speech signals are replayed with intermediate devices, such as high quality speakers, laptop speakers, and mobile phones.

Several countermeasures were approached by teams participated in BTAS 2016 challenge. Some of these countermeasures used Mel Frequency Cepstral Coefficients (MFCC) [26] fused with Inverse Mel Frequency Cepstral Coefficients (IMFCC) [27] using GMM [28, 29] as classifier, normalized perceptual linear prediction features [30] with Deep Neural Network (DNN), and Bi-directional Long Short Term Memory (BLSTM) as classifier, etc. The final evaluation performance is computed using Half Total Error Rate (HTER) which is the metric provided by the challenge organizers, ensuring fair comparison among all the participants [31].

In our recent study, we used Teager Energy Cepstral Coefficients (TECC) feature set [21] for classification of natural *vs.* replay speech. We observed that the Teager energy profiles are different for natural and replay speech signals. In [21], we linked the concept of reverberation along with Teager energy

10.21437/Odyssey.2020-43

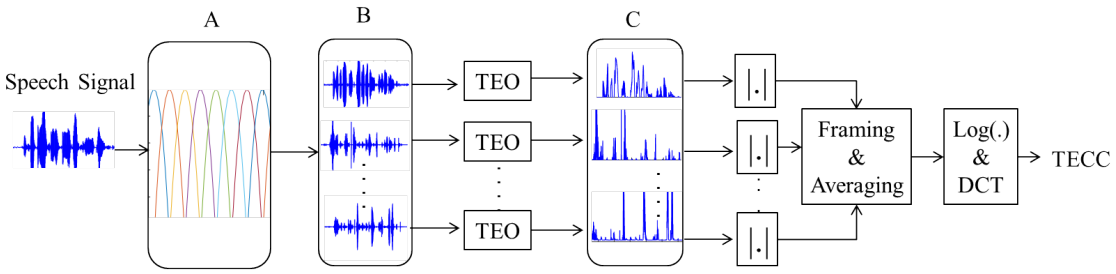

Figure 1: Block diagram of TECC feature extraction. A: Gabor filterbank, B: Narrowband filtered signals, and C: Teager energy profiles of each subband filtered signals. After [21].

profile to detect replay signal. These Teager energy profiles are useful to detect and classify the differences between natural and spoof speech signals. Hence, this paper is an extension of our earlier work exploring TECC feature set for SSD task. In this paper, we exploit Teager Energy-based feature set (i.e., Teager Energy Cepstral Coefficients (TECC)) for SSD task. In particular, we analyzed the Teager energy profiles of SS, VC, and replay speech signal. Furthermore, we compared the spectral energy densities obtained from the Teager energy *vs.* traditional spectrogram, and observed that Teager energy-based spectral patterns is capable to discriminate more compared to the traditional spectrogram between spoof, and natural speech signals.

2. Analysis of Teager Energy Profiles

An algorithm derived by Teager uses a nonlinear energy tracking operator [32]. For a monocomponent discrete-time signal, $x[n]$, Teager Energy Operator (TEO), $\Psi_d\{\cdot\}$, is defined as [32]:

$$E_n = \Psi_d\{x[n]\} = x^2[n] - x[n-1]x[n+1], \qquad (1)$$

where E_n gives the running estimate of signal's energy. Considering the speech signal, the TEO cannot be applied directly on the speech signal as it is the summation of multicomponent signals. Hence, the speech signal is bandpass filtered to obtain N number of narrowband filtered signal, and then the TEO is applied on the i^{th} narrowband filtered signal, i.e., $\Psi_d\{x_i[n]\}$.

The block diagram of Teager Energy Cepstral Coefficients (TECC) feature set is shown in Fig. 1. Originally, the TECC feature set was computed by first filtering the speech signal through a dense non-constant-Q Gammatone filterbank for robust speech recognition task [33], [34]. Here, the input speech signal is first given to the filterbank to obtain N=40 number of subband filtered signals [35], [36]. We have used linearly-spaced Gabor filterbank to have almost equal bandwidth to cover the entire frequency range [37, 38, 39]. Furthermore, these subband filtered signals are given to the TEO block to compute the TEO profile of each subband filtered signals. These TEO profiles 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) is then applied for energy compaction, and retained first few DCT coefficients to obtain TECC feature set, followed by their Δ and $\Delta\Delta$ feature vector to obtain higher-dimensional static plus dynamic feature vector.

3. Analysis of Spoof Speech Signals

We observed the Power Spectral Density (PSD) of natural (blue color), VC (pink color), and SS (red color) signal (from ASVspoof 2015 database) in Fig. 2.

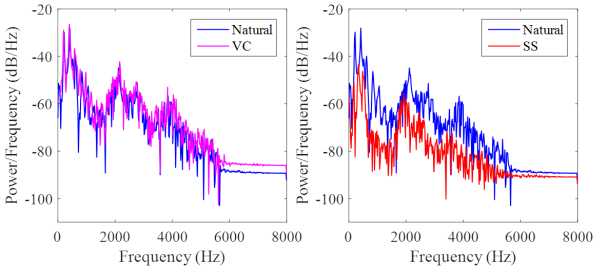

Figure 2: Power Spectral Density (PSD) for natural, and VC (left) and for natural and SS (right).

The PSD shows the stability of energy as a function of frequency, and energy (variations) are strong or they are weak at each frequency [40]. From Fig. 2(a), we can see very less difference between natural, and VC PSD plots they approximately overlap on each other, and have very less difference at higher frequency regions. On the other hand, the PSD obtained for natural and SS (as shown in Fig. 2(b)) shows very large difference almost for entire frequency regions.

Furthermore, the Teager energy profiles of the speech segment for natural (Panel I), VC (Panel II), and SS (Panel III) is analyzed as shown in Fig. 3. It is observed that the Teager energy traces obtained for a segment of natural speech signal have more energy, and more bumps are observed corresponding to the Glottal Closure Instant (GCI). Similar observation is found for segment of VC signal. However, the bumps around the GCI locations are very less compared to the Teager energy traces of natural signal. On the other hand, for the segment of SS signal, it can be observed that there are smooth bumps with very less fluctuations (indicating lesser energy modulations due to absence of natural speech production activities) in the instantaneous Teager energy traces compared to both natural, and VC bumps. This observation (highlighted with black box and arrows) is the key difference, and it helps to detect the VC and SS spoof signals from natural speech. In addition, we observed the difference in terms of spectral energies of Teager energy obtained from the output of the Gabor filterbank (as shown in Fig. 4). The spectral energy obtained from Teager energy for the natural speech preserves the formants and harmonics as shown in Fig. 4(a). Similar observation for VC signal is found with very less difference in the Teager energy (highlighted by the ovals) as shown in Fig. 4(b). The spectral energies obtained from Teager energy for SS signals shows the distorted and blurred energy

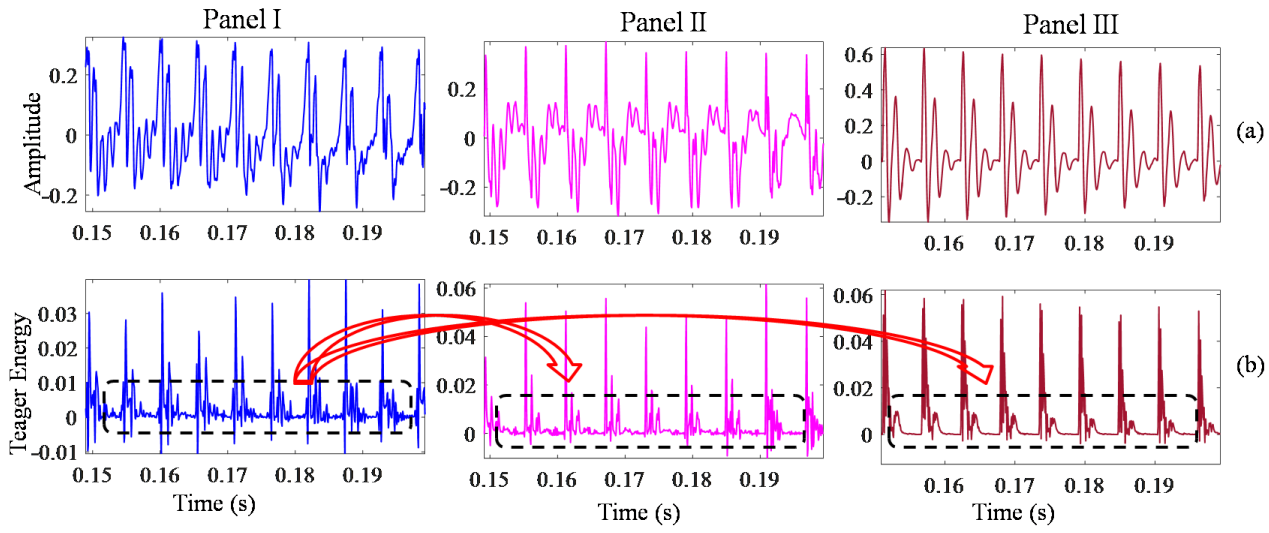

Figure 3: (a) Speech segment of natural (Panel I), VC (Panel II), and SS (Panel III) along with their corresponding Teager energy profiles in (b). Highlighted regions and arrows indicate change in Teager energy bumps (within two consecutive GCIs) for all the cases, in particular, for Panel III, the bumps in TEO profile are very smooth.

compared to the natural spectral Teager energy as shown in Fig. 4(c). We can see that there is loss in the energy and harmonics in the higher frequency regions (highlighted with box) in Fig. 4(c).

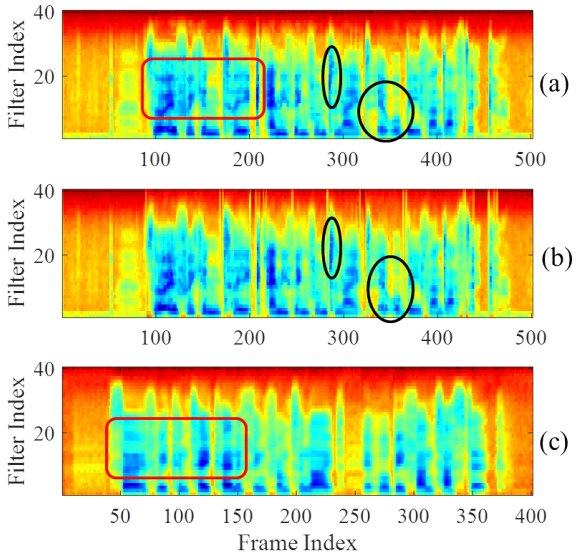

Figure 4: Comparison of Teager energy features for (a) natural, (b) VC, and (c) SS speech signal. Highlighted regions shows the difference between the natural *vs.* VC and SS.

Fig. 5 shows the (a) time-domain speech signal, spectral energies obtained from (b) Short-Time Fourier Transform (STFT), and (c) Teager energy-based method for all the speech signals (from BTAS 2016 competition dataset). The Panel I is for natural speech, and corresponding replay signals are shown in Panel II: Played back with Laptop, and Panel III: Played back with Laptop with high quality speaker, Panel IV and Panel V are corresponding synthesized, and voice converted speech signals that are played back with laptop and high quality speaker, respectively. For all the conditions in Fig. 5, it can be observed that the spectral energy density obtained from the Teager energy-based approach has high energy across entire frequency regions (because of linearly-spaced Gabor filterbank) as compared to the spectral energy density obtained from the traditional spectrogram. For natural speech signal, the formant frequencies have dark band color showing high energy portions of the speech signal. The shape of the dark bands shows the change from one sound unit to other w.r.t vocal tract shape. When we compare the energies of natural and replay speech signal, the replayed speech obtained with the high quality speaker device (Panel III) has similar pattern of energy and formant frequency band along with similar time-domain signal pattern. Whereas replay speech with normal quality device (Panel II) has distortions in the energies. For playback speech of machine-generated speech (i.e., Panel IV and Panel V), it can be observed that the spectral cues are not captured with traditional spectrogram Fig. 5(b), which is captured with the Teager energy approach and hence, helps to detect the natural *vs.* spoof speech signals.

The Teager energy profiles for a speech segment of natural (Panel I), replay laptop (Panel II), replay with HQ laptop (Panel III), SS with HQ laptop (Panel IV), and VC with HQ laptop (Panel V) are shown in Fig. 6. It can be observed that the Teager energy profiles obtained from various speech signals shows different energy profiles. However, Panel III shows similar pattern of Teager energy traces with natural speech segment, because replay signal is recorded, and replayed with HQ laptop device and hence, it is very similar to the natural counterpart and difficult to detect. It can also be observed from Table 4, the HTER for replay is better than the replay with HQ laptop. For Teager energy profiles of SS and VC, we can clearly observe the differences between the natural and replay speech signals. This is also strongly observed from our experimental results showing the lower HTER for SS and VC using HQ laptop as reported Table 4.

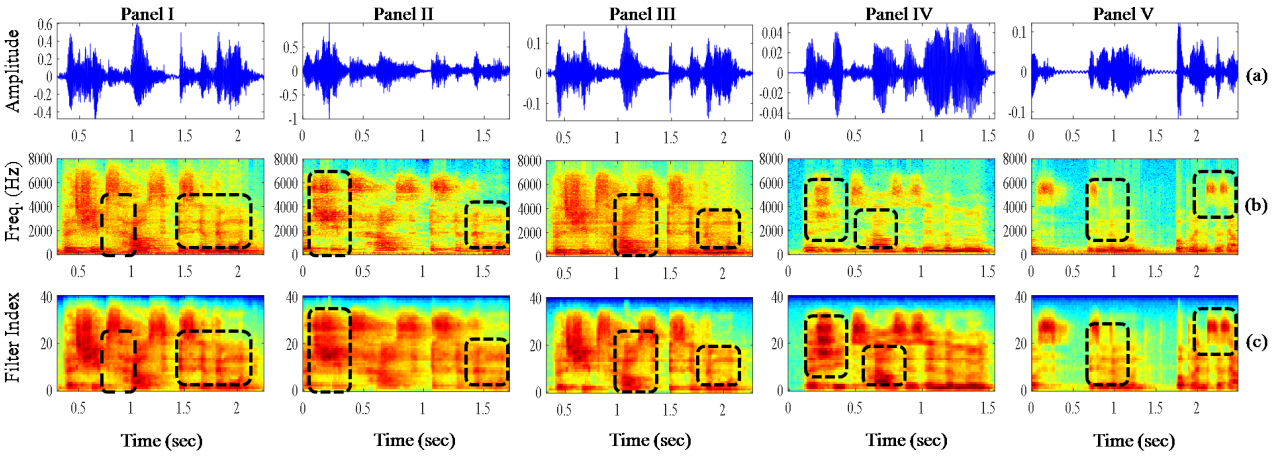

Figure 5: (a) Time-domain speech signal, spectral energy densities using (b) STFT spectrogram, and (c) Teager energy. Panel I: Natural, replay signals played back with Panel II: Laptop, and Panel III: Laptop HQ speaker, Panel IV: Speech Synthesis physical access HQ speaker, Panel V: Voice conversion physical access HQ speaker. Highlighted regions indicates the discriminative regions between the traditional spectrum and Teager energies.

.

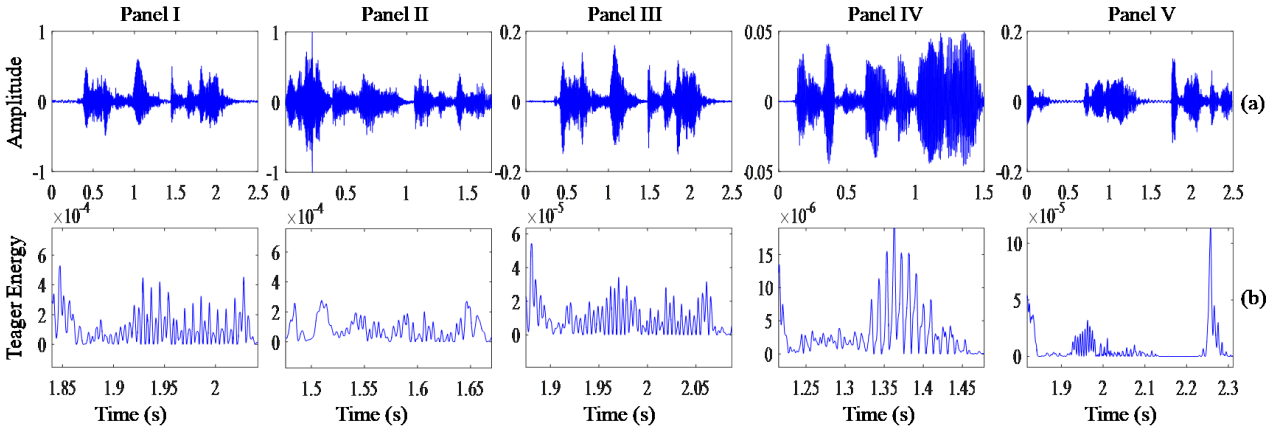

Figure 6: (a) Time-domain speech signal, and (b) Teager energy profiles. Panel I: Natural, replay signals played back with Panel II: Laptop, and Panel III: Laptop HQ speaker, Panel IV: Speech synthesis physical access HQ speaker, Panel V: Voice conversion physical access HQ speaker.

4. Experimental Setup

We used Gaussian Mixture Model (GMM) as classifier for modeling the classes corresponding to natural and spoofed speech utterances. Final scores are represented in terms of Log-Likelihood Ratio (LLR). The decision of the test speech being natural or spoofed is based on the scores of LLR:

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \quad (2)$$

where $P(X|H_0)$, and $P(X|H_1)$ are the likelihood scores of natural and spoofed speech trials with hypothesis H_0 and H_1, respectively. The score-level fusion is given by:

$$LLK_{fused} = \alpha LLK_{feature1} + (1-\alpha)LLK_{feature2}, \quad (3)$$

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

The performance evaluation metrics for BTAS 2016 database are considered according to the protocol used in the BTAS 2016 speaker anti-spoofing challenge. The results on the development data are reported in terms of EER and on the test data in terms of Half Total Error Rate (HTER). The evaluation of the replay attack systems was done based on the *false rejection rate* (FRR) and *false acceptance rate* (FAR), that in turn depends upon 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:

$$\theta_{dev} = \arg\min_{\theta} \frac{\text{FAR}_{dev}(\theta) + \text{FRR}_{dev}(\theta)}{2}, \quad (4)$$

$$\text{HTER}_{eval}(\theta) = \frac{\text{FAR}_{eval}(\theta_{dev}) + \text{FRR}_{eval}(\theta_{dev})}{2}. \quad (5)$$

5. Experimental Results

5.1. ASVspoof 2015 Database

The ASVspoof 2015 challenge database that was created for the ASV spoofing and countermeasure challenge, and it comprises of natural and spoof speech data [41]. Brief details of database are given in [7], [41]. The TECC feature set was extracted using 40 linearly-spaced Gabor filterbank with f_{min}=10 Hz, and f_{max}=8000 Hz. For each subband filtered signals, we obtain 40-dimensional (D) static features and further appended with their delta and double-delta coefficients resulting in 120-D feature vector to build the SSD system with 128 number of Gaussian mixtures in GMM classifier. We compared our results with other state-of-the-art features sets, such as Mel Frequency Cepstral Coefficients (MFCC) [42], Constant Q Cepstral Coefficients (CQCC) [11, 43], and Cochlear Filter Cepstral Coefficients-Instantaneous Frequency (CFCC-IF) [13].

5.1.1. ***Results on Development Set***

The results obtained in % Equal Error Rate (EER) of TECC feature set on development and evaluation sets are shown in Table 1. From the experimental results, it can be observed that on development set, the proposed feature set has much less % EER of 0.38 % compared to CFCC-IF and MFCC. However, the best performing feature set, i.e., CQCC gave lower % EER of 0.038 %. We further used score-level fusion of MFCC and TECC feature sets to obtain possible complementary information, and further reduce the % EER on both development and evaluation set. However, we could not obtain the reduced % EER.

Table 1: Comparison of results in % EER

Feature Set	Development	Evaluation
MFCC [42]	6.14	9.15
TECC	0.38	5.95
CFCC-IF [13]	2.29	1.211
CQCC [11]	**0.0381**	**0.255**
TECC+MFCC	0.38	6.41

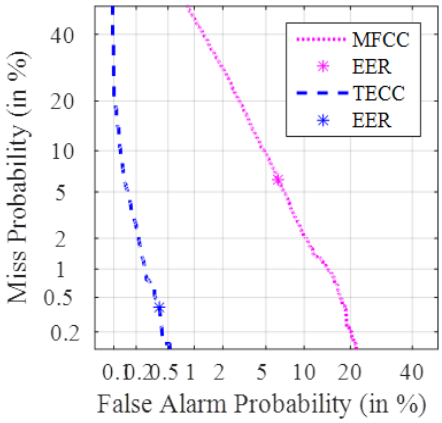

Figure 7: Individual DET curves of TECC and MFCC feature set on development dataset.

The performance is also shown in Fig. 7 by the Detection Error Trade-off (DET) curve on development set for MFCC and TECC feature sets. It can be observed from the DET curve that the operating points obtained from the score of MFCC have high miss probabilities and false alarm, whereas TECC feature set has a significantly lower false alarm and miss probabilities in the DET curve.

5.1.2. ***Results on Evaluation Set***

On evaluation set, the dataset is divided into two groups, namely, known (S1-S5) and unknown attacks (S6-S10). The unknown attacks were included during the challenge, which are not used in the training and development datasets. These unknown attacks are challenging to detect, in particular, the S10 attack which is developed with Unit Selection Synthesis (USS)-based approach. The detailed % EER of MFCC, TECC, CFCC-IF, and CQCC on both known and unknown attacks are reported in Table 2. It can be observed that for spoofing attacks (S1 to S9), for most of the cases, TECC feature set gave lower % EER compared to other state-of-the-art feature sets. For known attacks, the average % EER of TECC is 0.20 % and the average % EER for unknown attacks (S6-S9) is 0.161 % which is lower compared to other feature sets. The comparison in % EER from S1 to S9 spoofing algorithms are shown in Fig. 8. We can observe that the CFCC-IF feature set has higher % EER (green dotted line) compared to the CQCC and TECC feature set. For individual spoofing attacks of S7, S8, and S9, it can be observed that the % EER is equal to 0 % which is best performing system than the CQCC feature set. However, the TECC feature set fails to detect the S10 (USS) spoof speech signals resulting in higher % EER of 58.14 % that increases % EER for entire SSD task. This may be due to the fact that USS-based spoof contains concatenation of natural speech sound units results in similar bumps in TEO profile w.r.t nonlinearity in speech production and thus, creating a larger confusion during SS *vs.* natural SSD task.

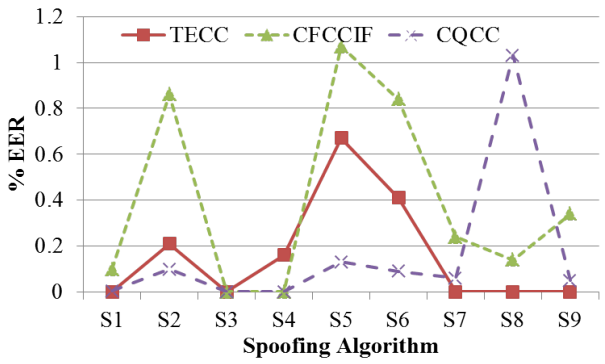

Figure 8: Comparison of S1-S9 spoofing algorithms in % EER of CFCC-IF (green line), CQCC (purple line), and TECC feature set (red line).

5.2. BTAS 2016 Database

The detailed statistics of the database is given in [44]. The organizers of the BTAS 2016 challenge provided a baseline system that uses the simple spectrogram-based ratio as features, and logistic regression as a classifier. In our experiments, the feature parameters used for TECC feature set is *120* - dimensional (*D*) (*40*-static +Δ+ $\Delta\Delta$). The TECC feature set is extracted from 40 linearly-scaled Gabor filterbank, and is compared with MFCC, CQCC feature set with the feature dimension of *39* -

Table 2: Results in % EER on Evaluation dataset for each spoofing attack. Both Known and unknown attacks. +:Score-level fusion

Feature Set	Known Attacks						Unknown Attacks						All Avg.	S1-S9 Avg.
	S1	S2	S3	S4	S5	Avg.	S6	S7	S8	S9	S10	Avg.		
MFCC	2.34	9.57	0.00	0.00	9.01	4.18	7.73	4.42	0.3	5.17	52.99	14.12	9.15	4.28
TECC	0.00	0.21	0.00	0.16	0.67	0.20	0.41	0.00	0.00	0.00	58.14	11.71	5.95	**0.161**
CFCC-IF	0.101	0.863	0.000	0.000	1.075	0.408	0.846	0.242	0.142	0.346	8.490	2.013	1.211	0.39
CQCC	0.005	0.106	0.000	0.000	0.130	0.048	0.098	0.064	1.033	0.053	1.065	0.462	0.255	0.163

dimensional (D) (*13*-static +Δ+$\Delta\Delta$), and *90*- dimensional (D) (*30* -static + Δ+ $\Delta\Delta$), respectively.

The results obtained in EER of TECC feature set on development and evaluation sets are shown in Table 3. We compared our results with the baseline system, MFCC, and CQCC feature set. From the experimental results, it can be observed that the TECC feature set has much more less EER of 2.25 % and 4.51 % on dev and eval set, respectively, compared to the baseline system, MFCC, and CQCC feature set.

Table 3: Equal Error Rate (EER) for BTAS 2016 Database

Subset	Baseline	MFCC	CQCC	TECC
Dev	5.91	3.66	3.05	**2.25**
Eval	-	7.59	18.86	**4.51**
Fusion with TECC				
Dev	-	**2.20**	2.25	-
Eval	-	**4.43**	4.50	-

We further used score-level fusion of MFCC and CQCC with TECC feature set to obtain possible complementary information, and reduce the % EER further on both development and evaluation set (as shown in Table 3). The score-level fusion reduced the % EER to 2.20 % with MFCC and TECC feature set (with fusion factor, α= 0.8) and with CQCC feature set it reduced to 2.31 % (with fusion factor α= 0.9). On the other hand, on evaluation set, the score-level fusion reduced only fusion of MFCC and TECC and gave % EER of 4.43 % (with fusion factor α= 0.9) whereas with CQCC feature, the EER did not reduce.

Table 4 shows the performance on evaluation set in % HTER on baseline system, MFCC, CQCC, and TECC feature set. It can be observed that TECC feature set gave lower % HTER compared to the other feature sets. Furthermore, we analyzed individual presentation attack as reported in Table 4. In the Table 4, 'SS' stands for speech synthesis, 'VC' stands for voice conversion, 'RE' stands for replay, 'LP' stands for laptop, 'PH1' is Samsung Galaxy s4 phone, 'PH2' is iphone 3gs, 'PH3' is iphone 6s, and 'HQ' stands for high quality speakers were used during replay. It can be observed that for all the attacks, we obtained lower % HTER with TECC feature set. However, for unknown attacks (highlighted with bold font), we obtained higher % HTER for all the feature sets which means degradation in the overall performance.

The histogram plots of log-likelihood scores obtained from Gaussian mixtures corresponding to (a) MFCC, (b) CQCC, and (c) TECC are shown in Figure 9 for development (Panel I) and evaluation set (Panel II), respectively. It can be observed that for TECC feature sets, the LLK scores of both natural and spoof are properly distributed resulting in less % EER as compared to the distribution corresponding to other feature sets on development set. Similar observation is found on evaluation set for MFCC, and TECC feature sets. From the Figure 9, we can observe a huge change in score distributions on development (i.e., -10 to 10) and evaluation (i.e., -80 to -10) sets for CQCC feature set. This in turn results in high % HTER for CQCC on evaluation set, as HTER depends on the threshold of development set (which is near to 0 (Figure 9 Panel I (b))).

Table 4: Individual Attack Results (in % HTER) for Eval Set

Attacks	Baseline	MFCC	CQCC	TECC
SS-LP-LP	2.87	10.82	50	2.39
SS-LP-HQ-LP	2.87	14.89	50	1.75
VC-LP-LP	3.58	4.05	50	1.43
VC-LP-HQ-LP	3.39	3.99	50	1.32
RE-LP-LP	17.02	9.40	50	1.77
RE-LP-HQ-LP	11.24	28.25	50	3.02
RE-PH1-LP	52.24	29.37	50	24.77
RE-PH2-LP	51.96	27.65	50	29.87
RE-PH2-PH3	51.56	38.85	50	50.17
RE-LPPH2-PH3	20.62	47.87	50	41.92
All together	6.87	6.89	50	**3.71**

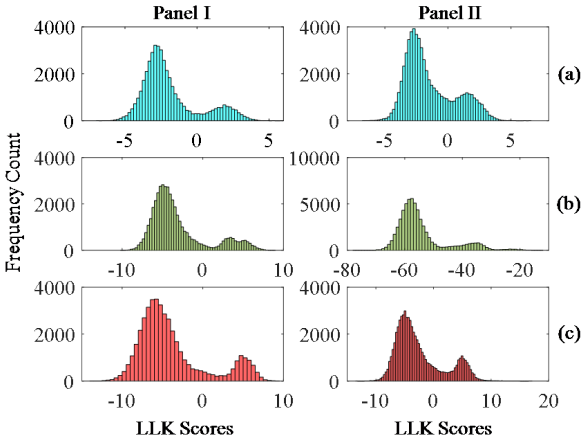

Figure 9: Histogram plots for Panel I: Development, and Panel II Evaluation set. (a) Score distribution of MFCC, (b) CQCC, and (c) TECC feature set.

The performance is also shown with DET curves for all the feature sets along with their best score-level fusion in Figure 10(a), and Figure 10(b). From Figure 10(a), it can be observed that for MFCC, and CQCC shows high miss probability and false alarm probability, respectively, which is not a good case for the voice biometric system. However, the TECC feature set along with score-level fusion of CQCC and TECC feature set shows the reduced miss probability and false alarm probability compared to the other feature sets. On the other hand, for evaluation set, the DET curve for all the feature sets have high probability with high false alarm rate which shows that the evaluation set is very challenging to develop a suitable countermeasure.

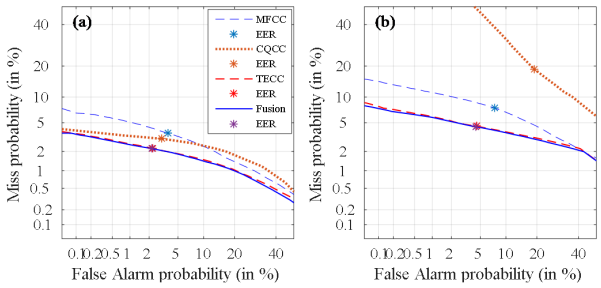

Figure 10: DET curve for (a) Development, and (b) Evaluation set.

6. Summary and Conclusions

In this paper, we investigated the significance of Teager energy profiles for SSD task, in particular, SS, VC, and replay speech signals. The Teager energy profiles of a narrowband filtered speech signal discriminates the spoof speech from the natural speech around the GCI locations. The bumps obtained around every GCI locations shows the key discrimination for natural, VC, SS, and replay signals. The Teager energy features have high energy for the natural speech compared to the spoof speech case. The results in EER with Teager energy-based feature set performed better on ASVspoof 2015 and BTAS 2016 challenge database than the other state-of-the-art feature sets. However, the TECC feature fails to detect the USS-based spoofing algorithm and unknown attack detection, in particular, the replay speech recorded with laptop HQ device.

Our future work will focus on the study of USS-based spoof detection. Furthermore, we studied the Teager energy profiles of the natural and presentation attack signals, and observed the changes in the Teager energy profiles. In particular, when the replay signal is generated using HQ laptop device, the Teager energy profiles are similar to the natural counterpart and thus, faces difficulty to detect the replay with laptop HQ compared to the other presentation attacks. We observed that although, we were quite successful in detecting certain kinds of presentation attacks, however, our system fails to detect unknown attack that are often expected in a practical scenarios. The negative result to detect S10 may find its relevance for significance of proposed TECC feature set for deeper analysis of speech excitation source characteristics, in particular, possible application in speaker recognition and recent efforts in Voice Privacy challenge.

7. References

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

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

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, pp. 1–19, 2020.

[10] T. Kinnunen, Z. Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4401–4404.

[11] 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.

[12] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2087–2091.

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

[14] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2082–2086.

[15] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2042–2046.

[16] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2092–2096.

[17] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Speech Communication*, vol. 81, pp. 30–41, 2016.

[18] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[19] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication, Elsevier*, vol. 85, pp. 43–52, 2016.

[20] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Odyssey 2016*, Bilbao, Spain, 2016, pp. 270–276.

[21] M. R. Kamble and H. A. Patil, "Analysis of reverberation via Teager energy features for replay spoof speech detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK*, pp. 2607–2611, 2019.

[22] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Adam's Mark Hotel Dallas, TX, USA: IEEE, 2010, pp. 1678–1681.

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

[24] 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.

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

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[27] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing*, vol. 4, no. 2, pp. 114–122, 2007.

[28] 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.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, pp. 1–38, 1977.

[30] H. Hermansky, "Perceptual linear predictive PLP analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[31] I. Chingovska, A. R. Dos Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2264–2276, 2014.

[32] 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.

[33] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *Ninth European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Portugal, 2005, pp. 3013–3016.

[34] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Processing Letters*, vol. 6, no. 10, pp. 259–261, 1999.

[35] 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.

[36] Maragos, Petros and Quatieri, Thomas F. and Kaiser, James F., "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.

[37] 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.

[38] 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.

[39] Madhu R. Kamble and Hemant A. Patil, "Novel amplitude weighted frequency modulation features for replay spoof detection," *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 185–189, Taipei, Taiwan 2018.

[40] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* 1^{st} *Edition*. Pearson Education India, 2006.

[41] 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.

[42] 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.

[43] 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.

[44] 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* 8^{th} *International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Niagara Falls, New York, USA, 2016, pp. 1–6.