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

Novel Variable Length Teager Energy Profiles for Replay Spoof Detection

Madhu R. Kamble and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of
Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India
{madhu_kamble, hemant_patil}@daiict.ac.in

Abstract

Replay attacks are developed in order to get fraudulent access of an Automatic Speaker Verification (ASV) system. This attack requires only recording and playback devices. The replay speech gets affected by the use of quality of intermediate devices, and the level of noise present in the acoustic environment. In this paper, we propose Variable length Teager Energy Cepstral Coefficients (VTECC) for replay Spoof Speech Detection (SSD) task. Varying the Dependency Index in Variable length Teager Energy Operator (VTEO) changes the performance of SSD system. The Teager energy profiles and the spectral energy densities obtained show the discrimination information for different DIs. With DI=5, we got reduced % Equal Error Rate (EER) of 6.52 % and 11.93 % on development and evaluation set, respectively, on ASVspoof 2017 version 2.0 challenge database. We further used score-level fusion of baseline system (Constant Q Cepstral Coefficients (CQCC) feature set) and VTECC and reduced the % EER to 5.85 % and 10.94 % on development and evaluation set, respectively. Furthermore, for evaluation set, we investigate the performance on different Replay Configurations (RC). For all the levels of threats, the proposed feature set performed better compared to the other feature sets.

keywords: Automatic Speaker Verification (ASV), Spoofing, Replay, Variable length Teager Energy Operator (VTEO), Replay Configurations (RC).

1. Introduction

The growing technological development and improvements in various biometrics leads to various spoofing attacks [1, 2]. The Automatic Speaker Verification (ASV) systems verify the claimed speaker's identity from their voice [3, 4]. Spoofing refers to attacks where a fraudsters attempts to gain access of the system by masquerading as an enrolled person in the ASV system [5, 6]. The present ASV system is susceptible to various spoofing attacks, such as speech synthesis (SS), voice conversion (VC), replay, impersonation, and twins [5, 7, 8, 6]. Since from last few years the research in spoofing and countermeasure has attracted significant attention from the industry, academics, forensics, government projects, etc.

Replay attack is one of the most accessible spoofing attack [9]. The attacker replays a pre-recorded voice from the target speaker to the system to gain access [10, 11, 12]. The ASVspoof 2017 challenge provided a common platform with standard corpora, protocol, and metrics focusing exclusively on replay spoofing attacks [13]. The organizers of the challenge also provided the baseline system that includes Constant Q Cepstral Coefficients (CQCC) as a front-end feature set and Gaussian Mixture Model (GMM) as back-end classifier [13], [14]. However, in the database organizers came across some anomalies files and replaced the database with a few modifications that results in the modified version of database, i.e., ASVspoof 2017 challenge version 2.0 database.

The research on the non-publicly available databases were started long way back [15, 12, 16]. The spectral peak mapping method was proposed as a countermeasure to detect the replay attack on a remote telephone interaction [15]. Replay attacks with far-field recordings were addressed in [12]. The acoustic features, such as Rectangular Filter Cepstral Coefficients (RFCC), Subband Spectral Centroid Magnitude Coefficients (SCMC), Subband Spectral Centroid Frequency Coefficients (SCFC), and Subband Spectral Flux Coefficients (SSFC) were used to detect replay speech and found that the SCMC followed by feature normalization method performed better than various other acoustic features [17]. Several other features were also used for ASVspoof 2017 challenge database [18, 19, 20, 21, 22, 23], etc.

In our earlier studies, we explored the Teager Energy Operator (TEO) and Energy Separation Algorithm (ESA)-based feature for Spoof Speech Detection (SSD) task in [24, 25, 26, 27, 21, 22, 20]. In this paper, we are exploring the variable length version of TEO for replay SSD task with varying the Dependency Index (DI) also known as *lag parameter*. This lag parameter captures the hidden dependencies of the narrowband filtered Teager energy profiles and thus, helps to classify the replay signals from its natural counterparts.

The organization of rest of the paper is as follows. The analysis along with the brief details of feature sets used is presented in Section 2. Brief details of the database, and experimental setup are presented in Section 3. Section 4 presents the experimental results on ASVspoof 2017 challenge version 2.0 database along with analysis of different replay configuration. Finally, Section 5 concludes the paper along with the future research directions.

2. Variable length TEO

The TEO tracks the running estimate of instantaneous energy fluctuations of the narrowband filtered speech signal. The Teager energy profile obtained from the narrowband filtered signals can approximately estimate the squared product of IA ($a_i[n]$), and IF ($\Omega_i[n]$) for the i^{th} subband filtered signal is given as [28, 29]:

$$\Psi_d\{x[n]\} = x_i^2[n] - x_i[n-1]x_i[n+1] \approx a_i^2[n]\Omega_i^2[n]. \quad (1)$$

Variable length Teager Energy Operator (VTEO) is the modified version of the traditional TEO method [30]. TEO involves nonlinear operations on the signal, i.e, square of current sample and multiplication of previous and next sample, i.e., $x(n-1)$ and $x(n+1)$, respectively. The key motivation for VTEO is the speech signal carries dependencies (local

10.21437/Odyssey.2020-21

vs. distant) in the sequence of samples of speech signal. Thus, instead of considering only immediate past $x[n-1]$ and immediate future $x[n+1]$, VTEO considering k^{th} past and k^{th} future samples. In VTEO algorithm, the number of samples incorporated in energy estimation can be varied up to k past, and k future samples, i.e., $x(n-k)$ and $x(n+k)$, instead of only two adjacent samples as in TEO [31]. VTEO gives flexibility to select these samples to estimate the running estimate of energy required to generate the signal [32]. VTEO gives us a good measure of the energy of the oscillating signal, when the sampling rate of the signal is greater than $8i$ times the frequency of oscillation in the signal [31]. VTEO brings out *hidden* dependencies and dynamics of the signal [31]. For discrete-time signal, $x[n] = Acos(\omega n + \phi)$, the samples of the same signal shifted in time by index k, w.r.t present sample, can be expressed with an assumption for $k¿n$, $x(n-k) = 0$ as:

$$x(n+k) = Acos(\omega(n+k)+\phi) \tag{2}$$

$$x(n-k) = Acos(\omega(n-k)+\phi) \tag{3}$$

When we multiply above equations we obtain,

$$x(n+k)x(n-k) = A^2 cos(\omega(n+k)+\phi)cos(\omega(n-k)+\phi) \tag{4}$$

$$x(n+k)x(n-k) = [Acos(\omega n+\phi)]^2 - A^2 sin^2 i\omega. \tag{5}$$

On high sampling rates it result to VTEO and is given as Eq. (6):

$$E_n = \{\Psi_{DI}\{x(n)\}\} = x^2(n) - x(n-k)x(n+k) \approx k^2A^2\omega^2, \tag{6}$$

where $k^2A^2\omega^2$ is instantaneous estimate of signal's energy multiplied by k^2, and referred to as VTEO for the dependency index (DI), k, which is expected to give running estimate of signal's energy [32, 33].

The VTEO has the superior property w.r.t. localization and tracking instantaneous energy of a narrowband signal. It also brings out the *hidden* dependencies and dynamics of the signal w.r.t. distantly located speech samples than only immediate adjacent samples (as in case of traditional TEO).

2.1. Feature Extraction Process

The block diagram of Variable length Teager Energy Cepstral Coefficients (VTECC) feature set is shown in Figure 1. VTECC is an extension of our recent study reported in [20, 34]. VTECC is found to perform better for SSD task, synthetic and converted speech (SS and VC) signal as per our recent work done on the ASVspoof 2015 challenge database [34]. The VTECC was computed by first filtering the speech signal through a dense non-constant-Q Gammatone filterbank for robust speech recognition in [35, 36]. The input speech signal is given to the filterbank to obtain N number of subband signals [37, 28]. We have used linearly-spaced Gabor filterbank to have almost equal bandwidth to cover the entire frequency range [21, 22, 27]. Furthermore, these subband filtered signals are given as input to the TEO block to compute the energy 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 VTECC feature set, followed by their Δ and $\Delta\Delta$ feature vector to obtain higher-dimensional static plus dynamic feature vector. From the earlier studies on replay SSD task, we found that the higher frequency regions are more useful along with Cepstral Mean Normalization (CMN) technique. Hence, VTECC feature set is extracted using pre-emphasis filter and CMN technique [21, 22].

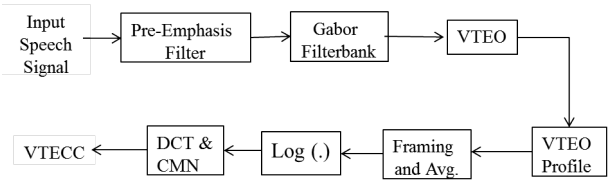

Figure 1: Block diagram of proposed variable length Teager energy cepstral coefficients (VTECC). After [20].

In our earlier study [20], we tried to link the concept of reverberation with replay SSD task, as the replay signal are recorded and played back, where the reverberation exist. In Figure 2, the synthetic sinusoidal signals (Panel I) are shown along with their corresponding TEO profiles (Panel II). Figure 2(a) show the damped sinusoidal signal with equal amplitude of impulse and Figure 2(b) show the damped sinusoidal signal with decrease in amplitudes of the impulse. Whereas, Figure 2(c) show the variations in the amplitude of the damped sinusoidal signal. It can be observed from their corresponding TEO profiles in Panel II that for each case the TEO show impulse-like energies. In particular, if the amplitude of the signal is constant the TEO profiles are also constant in terms of its amplitude, and if the amplitude of signal varies (as in case on Panel I (b and c)) the corresponding TEO profiles also varies (highlighted by the box and oval shapes).

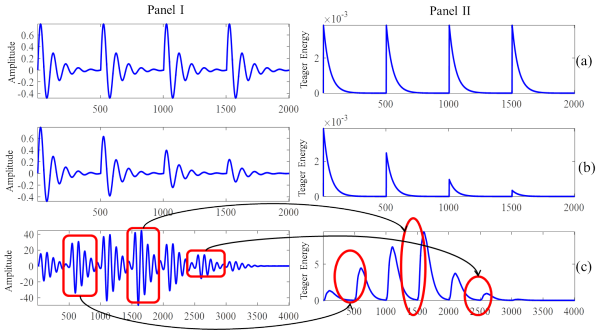

Figure 2: Panel I: Synthetic sinusoidal signals with (a) same, (b) decreasing, (c) varying sinusoidal signals along with their corresponding Teager energy profiles in Panel II.

The TEO profiles show high energy pulses around the Glottal Closure Instant (GCI), because of impulse-like excitation to vocal tract system and this sudden glottal closure produces high energy and thus, TEO produces high energy around these regions [38]. Along with high Teager energy pulses, the bumps are observed around the energy pulses, indicating significant contribution of nonlinear effects during the speech production process [38]. This nonlinear effect is observed for real speech signal as shown in Figure 3, in particular, for natural (Figure 3(a)) and its corresponding replay speech signal (Figure 3(b)). When compared to the synthetic signal as shown in Figure 2 the nonlinearities around the GCI locations are missing and hence, the natural speech confirms the capability of Teager energy to

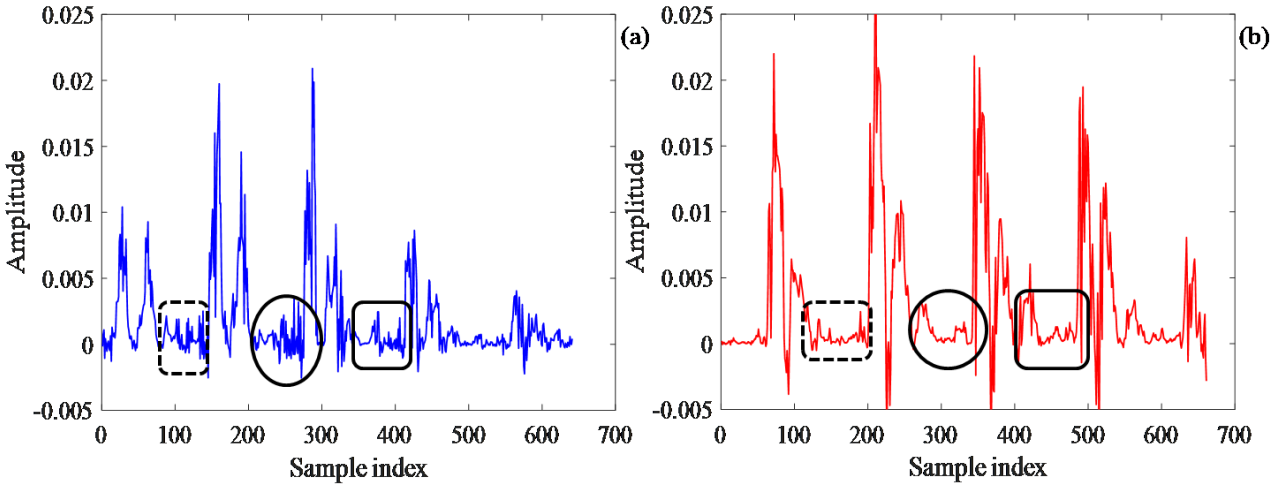

Figure 3: Teager energy profiles for (a) natural and (b) replay speech segment. Highlighted regions shows the contribution of nonlinear effects during speech production process which is not observed for synthetic case.

represent characteristics of airflow pattern during natural speech production.

We observed the Teager energy traces of the speech segment considered for natural (blue line) and replay (red line) as shown in Figure 4. We can see that for the segment of replay speech very high (impulse-like) energy traces are obtained when compared to the segment of natural speech. In addition, we also observed the PSD for Teager energy traces of natural and replay speech segment as shown in Figure 5. The variation at each frequency component for Teager energy traces of replay segment (red line) are more smooth compared to that of Teager energy traces of natural segment (blue line).

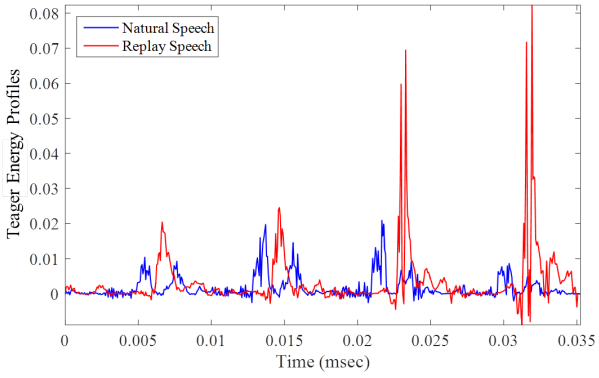

Figure 4: Teager energy traces of the natural (blue line) and replay (red line) speech segment.

2.2. Analysis of Variable length Teager Energy Profiles

The VTEO profiles corresponding to DI= 1 to 10 are shown in Figure 6. The blue line corresponds to natural Teager energy profiles, and red line to replay speech signals. For the initial DI's, i.e., from 1 to 2 for replay signal we cannot see the profiles clearly they are all merged around the glottal closure instant's (GCI's). After DI=2 the replay signal profiles start to show the Teager energy profiles similar to the natural signal. Later after DI=6 more fluctuations and bumps are observed in replay signal whereas, it is reduced for the case of natural signal as we increase the DI after 6. According to the results shown in experimental result section with DI=5 the replay signals are detected and classified well compared to other DI's.

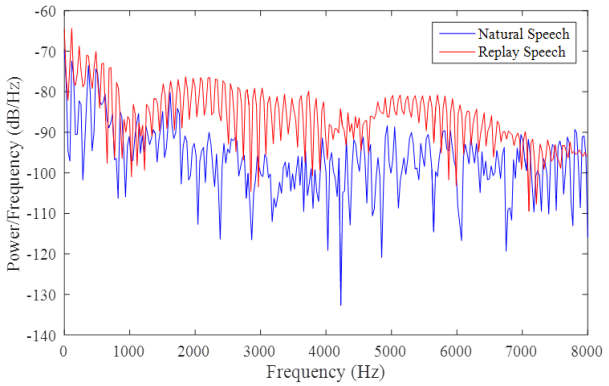

Figure 5: Power Spectral Density (PSD) of Teager energy traces of the natural and replay speech segment

2.3. Spectral Energies of Variable length Teager Energy

Figure 7 show the spectral energy corresponding to each DI obtained from Variable length Teager energy. The spectral energies here is shown for the natural speech signal. It can be observed from the Figure 7 that with every DI we find some differences corresponding to the first DI (shown by highlighted circles). With DI=5, we observe more spectral energy differences in lower as well as in higher frequency regions. This spectral energy changes corresponding to other DI helps to detect and classify it from the natural signal. This can also be observed from the results obtained from all the DI's reported in Section 4 were we obtained relatively lower % EER at DI=5.

3. Experimental Setup

The experiments were performed on the ASV Spoof 2017 challenge version 2.0 database and the detailed statistics of the database is given in [39]. Following state-of-the-art techniques were explored for the replay SSD task.

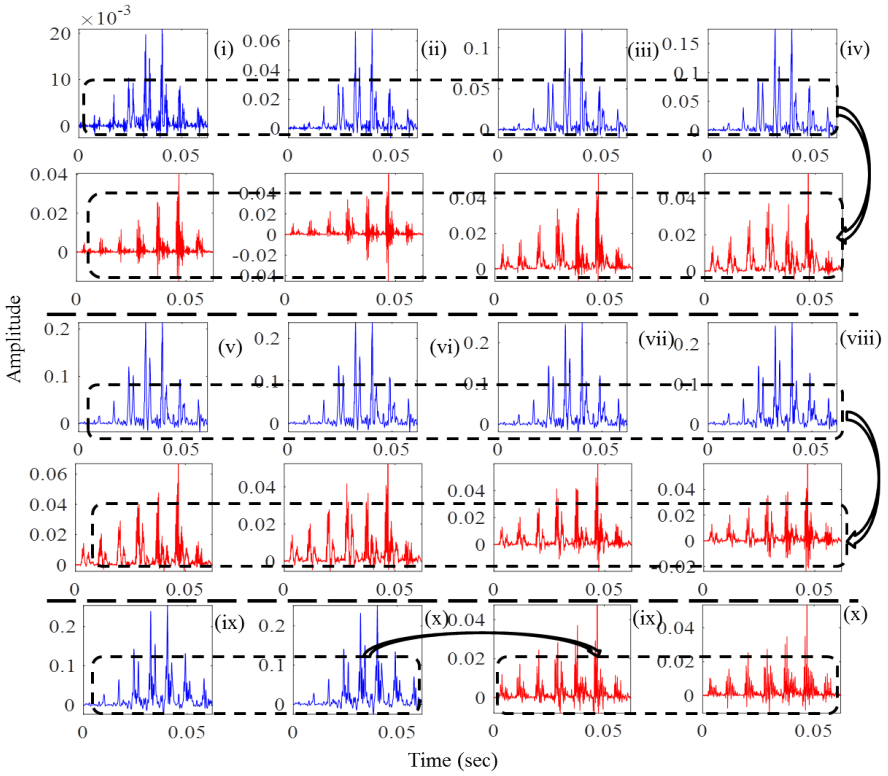

Figure 6: Teager energy profiles with varying the DI from 1 to 10. Blue and red colored corresponds to Teager energy profiles of natural and replay spoof signal, respectively. Highlighted regions show the difference in Teager energy profiles.

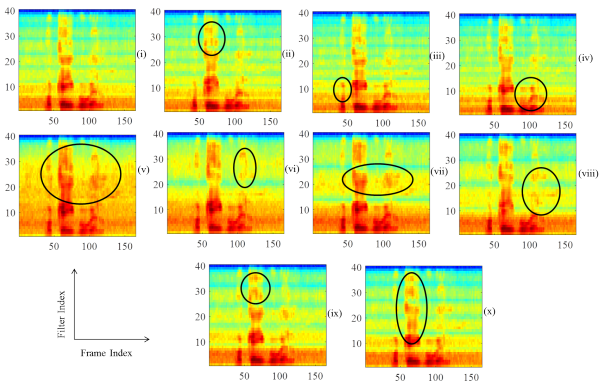

Figure 7: Spectral energy densities obtained from variable length Teager energy with varying the DI=1 to 10 for (i) to (x), respectively from natural speech via variable length Teager energy operator. Highlighted circles show the differences in spectral energies.

- **Baseline System**: The CQCC features are extracted with $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8 kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \approx 15Hz$. The number of bins per octave B is set to 96. Features extracted with 30 DCT static coefficients (with log-energy), resulting in total 90-D (static+delta+delta) feature vector [14, 40].

- **LFCC**: The LFCC feature set is extracted with *20* DCT static coefficients, resulting in total *60*-D feature vector (including *20*-Δ and *20*-$\Delta\Delta$) [41].

- **MFCC**: The MFCC feature set is extracted from 40 Mel filterbank and retained *13* DCT static coefficients, resulting in total *39*-D feature vector (including *13*-Δ and *13*-$\Delta\Delta$) [20].

  While extracting MFCC or LFCC feature sets, the speech signal is windowed and DFT is computed for each frame to get the Short-Time Fourier Transform (STFT), $X(n, \omega_k)$. The energy in STFT is weighted by each Mel scale filter frequency response, $V_l(\omega)$, to get the l^{th} energy coefficient [42], i.e.,

$$E_{mel}(n, l) = \frac{1}{A_l} \sum_{k=L_l}^{U_l} |V_l(\omega_k) X(n, \omega_k)|^2. \quad (7)$$

  The real cepstrum C_{mel} associated with the $E_{mel}(n, l)$ is referred to as MFCC:

$$C_{mel}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} log(E_{mel}(n, l)) cos(\frac{2\pi}{R} lm), \quad (8)$$

  where R is the number of subband filters. The transformation in eq. (8) is also known as Discrete Cosine Transform (DCT). Both MFCC, and LFCC use similar algorithm for feature extraction except the type of frequency response used to obtain the weighted sum from the spectrum. In general, Mel scale gives more significance (resolution) to the lower frequency regions, and less significance to the higher frequency regions [43].

This arrangement suggests that the MFCC fails to extract effective spectral characteristics at the high frequency regions. Both MFCC and LFCC feature sets use triangular-shaped filters in order to obtain the subband filtered components.

- **VTECC**: The VTECC feature set was extracted using 40 linearly-spaced Gabor filterbank with f_{min}=10 Hz, and f_{max}=8000 Hz [20]. For each subband filtered signals, we obtain *40*-dimensional (D) static features appended along with their Δ and $\Delta\Delta$ coefficients resulting in *120*-D feature vector to build the SSD system.

We have used GMM classifier for modeling the classes corresponding to natural and spoofed speech utterances. GMM is a more popular and well known classification technique widely used in signal processing and pattern recognition literature [44]. GMM is a generative model that represent each class as a weighted sum of *M* multivariate Gaussians and it is given by:

$$p(x|\lambda) = \sum_{k=1}^{M} w_k p_k(x), \quad (9)$$

where w_k is the k^{th} mixture weight, and $p_k(x)$ is a *D*-variate Gaussian probability density function with mean vector μ_i and covariance matrix Σ_i. The model parameter is defined by λ. 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:

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \quad (10)$$

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:

$$LLR_{fused} = \alpha LLR_{feature1} + (1-\alpha)LLR_{feature2}, \quad (11)$$

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

4. Experimental Results

4.1. Results with varying Dependency Index (DI)

The results with varying the DI from 1 to 10 on development set of the proposed VTECC feature set is shown in Figure 8. The VTECC feature set obtained an EER of 9.55 % with DI=1, whereas with DI=5 the EER is reduced to 6.52 % which is relatively least % EER among all the DIs. This is because of the spectral energies obtained with DI=5 has more energy as observed and discussed in Figure 7. Hence, for further set of experiments, we considered DI=5 for VETO computation.

We compared our results with state-of-the-art features, such as CQCC, LFCC, and MFCC. The results obtained from these feature sets for both development and evaluation set are reported in Table 1. Here, the CQCC feature set which is baseline system is extracted in cepstral-domain whereas in actual baseline system, the organizers used log-energy coefficients. The histogram plots of log-likelihood scores obtained from Gaussian mixtures corresponding to (a) CQCC, (b) LFCC, (c) MFCC, and (d) VTECC are shown in Figure 9 for development set. It can be observed that with VTECC feature set, the LLR scores of both natural and replay are distributed more resulting in lower % EER as compared to the distribution obtained from other feature set on development set.

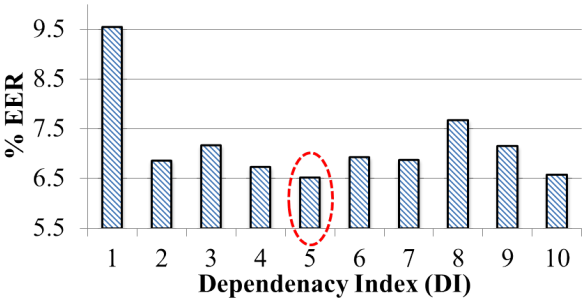

Figure 8: Bar graph plot with varying the DI on development set. Highlighted circle indicates the least % EER.

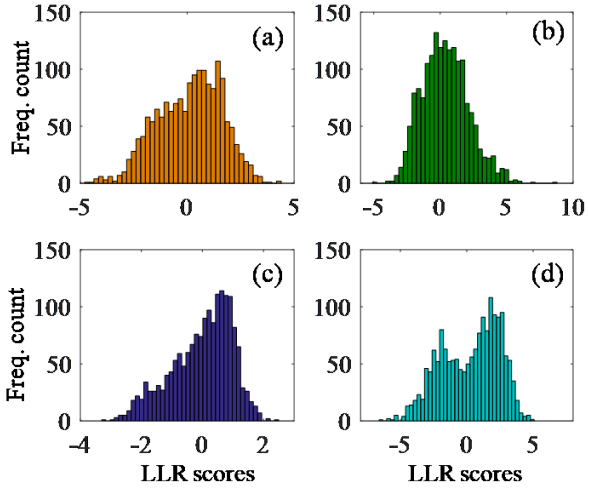

Figure 9: Histogram plots of (a) CQCC, (b) LFCC, (c) MFCC, and (d) VTECC feature set on development set..

4.2. Results with Score-level Fusion

In addition to the individual performance of the feature sets, we further performed the score-level fusion in order to investigate the possible complementary information of the feature sets, and reduce the % EER further. The comparison of all the feature sets along with their score-level fusion of VTECC feature set with CQCC, LFCC, and MFCC is shown in Table 1. It can be observed that the individual performances on development and evaluation set has higher % EER compared to the VTECC feature set. The % EER is further reduced when the CQCC and VTECC feature sets are fused at score-level reducing the % EER to 5.85 % and 10.94 % at fusion factor α=0.6 and α=0.8 on development and evaluation set, respectively.

Table 1: Final Results on Dev and Eval Set

Feature Set	Dev	Eval
CQCC (Baseline)	12.75	18.97
LFCC	10.31	15.73
MFCC	23.80	26.62
VTECC	6.52	11.93
CQCC+VTECC	**5.85**	**10.94**
LFCC+VTECC	6.52	11.93
MFCC+VTECC	6.52	11.67

Proposed VTECC is computed with DI=5

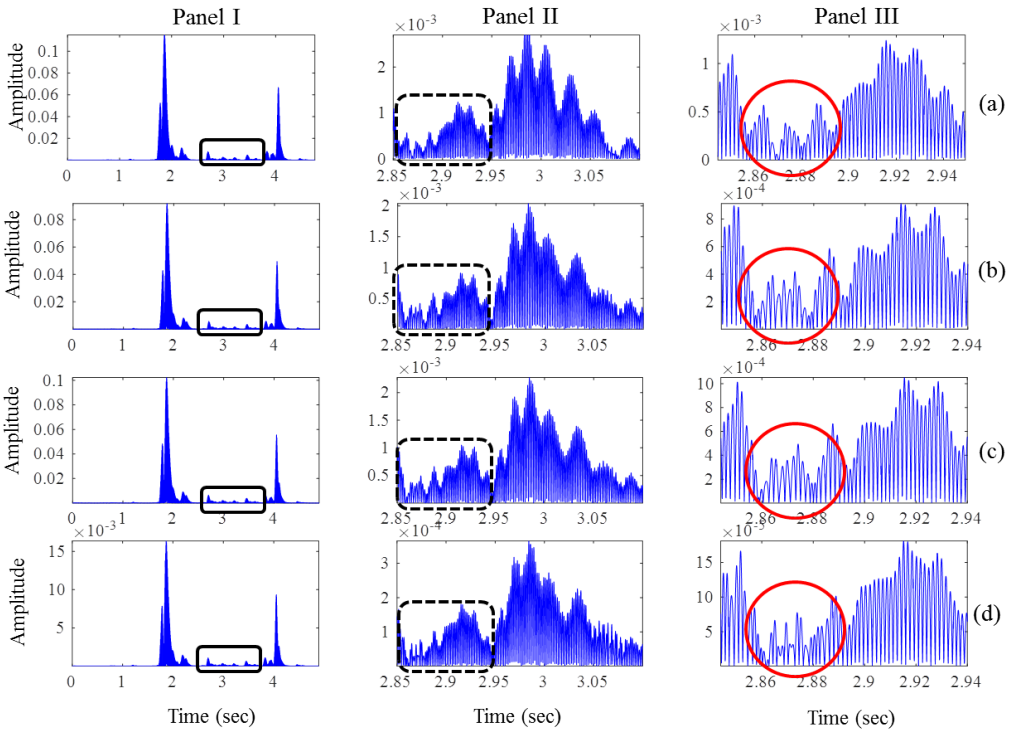

Figure 10: Panel I: Bandpass filtered signal around first formant. Panel II: zoomed version of the above signal panel I. Panel III: Temporal fine structure of Panel II with different time scale (a) natural, and replay with (b) perfect, (c) high and (d) low quality devices

The performance is also shown in Figure 11 with DET curves for all the feature sets along with their best score-level fusion on development and evaluation set, respectively. From Figure 11(a), it can be observed that for MFCC, CQCC, and LFCC show high miss probability and false alarm probability which is not a good case for the voice biometric system. However, the VTECC feature set along with score-level fusion with CQCC and MFCC feature set show the reduced miss probability and false alarm probability compared to the other feature sets. On the other hand, for evaluation set, the DET curves for all the feature sets have high probability with high false alarm rate this show that the evaluation set is challenging for given SSD taask.

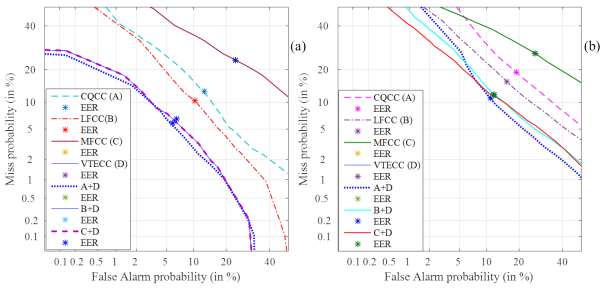

Figure 11: Individual DET curves of different feature sets (a) development, and (b) evaluation set.

4.3. Results on Replay Configurations (RC)

The physical significance in terms of temporal modulations at different time scale is analyzed in Figure 10. The time-domain subband filtered signal around 1st formant frequency is shown in Figure 10(Panel I) for (a) natural, and replay with (b) perfect, (c) high, and (d) low quality devices. The slow temporal modulations of a speech signal roughly correlates with the different syllabic segments. For natural speech, slow temporal modulations results in smooth amplitude envelope as shown in Figure 10(a) (in Panel II). The higher peaks in the fast temporal modulations (which are also known as Temporal Fine Structure (TFS)) as shown in Panel III of Figure 10(a) represents the harmonic structure of the speech signal. However, this observation is missing for the replay speech (Panel II) of Figure 10(b-d). The slow temporal modulations for replay speech are having distorted amplitude envelope (Panel II) of Figure 10(b). While the fast temporal modulations do not represents the harmonic structure Figure 10(b-d) of Panel III. It can be observed from the slow temporal modulations of replay speech that the variations are very less. On the other hand the fast temporal modulations indeed show the differences for different quality of intermediate devices varying from the perfect, high, and low. The perfect and high quality device have the similar pattern of fast temporal modulations however, this analysis could be very useful for the speech signal when recorded in low quality devices (as observed in Panel III of Figure 10(d)).

The level of noise in acoustic environment, playback, and recording device are assumed to be inversely proportional to the threat for ASV system pose [39]. The Replay Configurations consists of acoustic environment, playback and recording devices, respectively. These RCs are further classified into three different threat levels, namely, low, medium, and high. Different environments have the variations with the levels of additive ambient, convolutive, and reverberation noise. According to the different RC, the % EER of VTECC feature set along

with CQCC, LFCC, and MFCC are shown in Figure 12. The least % EER for all the RCs are obtained with the proposed VTECC feature set. It can be observed that for all the RC the % EER for MFCC feature set are too high compared to the LFCC and CQCC feature sets. The high-level threats are difficult to detect due to use of professional audio equipment, such as active studio monitors, studio headphones, etc. to produce replay samples [39].

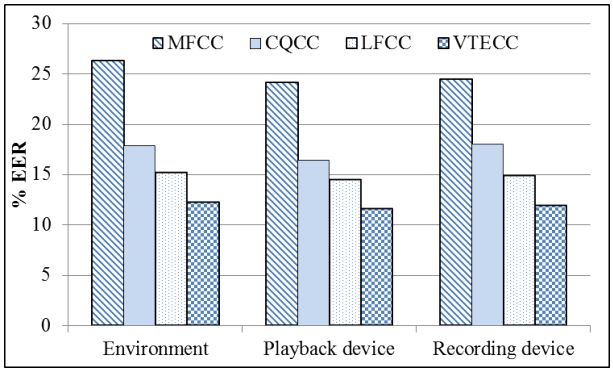

Figure 12: Bar graph representation for different replay configurations, i.e., acoustic environment, playback, and recording device (results in % EER).

5. Summary and Conclusions

In this paper, we investigated the significance of Variable length Teager energy profiles for replay SSD task. The variable length Teager energy profiles are obtained from the linearly-spaced Gabor filterbank that discriminates the replay speech from the natural speech around the GCI locations. In particular, the DI in TEO have the advantage (over the traditional TEO) of having superior localization and tracking instantaneous energy of the narrowband VTEO profile. With change in the DI the spectral energies of signal also changes. We also observed the changes in the VTEO profiles for DI=1 to 10, and found that as we increase the DI, the bumps around impulse-like peaks increases which is the key difference between natural and replay speech signal. With VTECC, we get lower % EER for replay SSD task and also gave lower % EER for all replay configurations compared to the other state-of-the-art feature sets. One of the limitation of this work is the performance of SSD system that is optimized w.r.t Dependency Index (DI). Our future research efforts are directed towards performance on recent ASVspoof 2019 real PA and ASVspoof 2019 challenge database.

6. Acknowledgments

The authors would like to thank University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF) and authorities of DA-IICT Gandhinagar for their kind support and co-operation to carry out this research work. We also thank the organizers of ASVspoof 2017 challenge.

7. References

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

[2] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, vol. 79, no. 1, pp. 80–105, 2016.

[3] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of Biometric Anti-Spoofing*, Springer, 2014, pp. 125–146.

[4] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, "Handbook of biometric anti-spoofing: Presentation attack detection," *Advances in Computer Vision and Pattern Recognition*, Springer, 2018.

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

[6] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (ASV) system," *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, pp. 1047–1053, Hawaii, USA, 2018.

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

[8] 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.

[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, Cambridge University Press*, vol. 9, pp. 1–18, 2020.

[10] J. Lindberg, M. Blomberg *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques," in *EUROSPEECH*, vol. 99, Budapest, Hungary, 1999, pp. 1211–1214.

[11] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, Vigo, Spain, 2010, pp. 131–134.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

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

[19] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 32–36.

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

[21] 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.

[22] 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.

[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] 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.

[25] 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.

[26] 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.

[27] M. R. Kamble and H. 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.

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

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

[30] 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.

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

[32] H. A. Patil and K. K. Parhi, "Novel variable length Teager energy based features for person recognition from their hum," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Adam's Mark Hotel Dallas, TX, USA: IEEE, 2010, pp. 4526–4529.

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

[34] M. R. Kamble, A. K. S. Pulikonda, S. K. Maddala, and H. A. Patil, "Analysis of Teager energy profiles for spoof speech detection," *to appear in Odyssey The Speaker and Language Recognition Workshop*, Tokyo, Japan, 2020.

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

[36] 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.

[37] 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.

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

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

[40] 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.

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

[42] 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.

[43] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[44] 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.