ANALYSIS OF REVERBERATION VIA TEAGER ENERGY FEATURES FOR REPLAY SPOOF SPEECH DETECTION

Madhu R. Kamble and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information
and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India.
E-mail: {madhu_kamble and hemant_patil}@daiict.ac.in

ABSTRACT

The Automatic Speaker Verification (ASV) systems are vulnerable to spoofing attacks. Detecting replay attack is the challenging Spoof Speech Detection (SSD) task, as several factors are involved during replay mechanism. Hence, it is important to analyze these factors for effective SSD task. This paper introduces the analysis of the replay speech focusing only on the effect of reverberation on the replay speech. The reverberation introduces delay and change in amplitude producing close copies of natural signal that makes natural components inseparable from the replay components and hence, fails to classify the replay speech signal. To that effect, we propose use of Teager Energy Operator (TEO) to compute running estimate of subband energies for replay *vs.* natural signal. These subband energies are mapped to cepstral-domain to get proposed Teager Energy Cepstral Coefficients (TECC) for replay SSD task. With the TECC feature set, we analyzed the individual performance for all the Relay Configurations (RC) with Gaussian Mixture Model (GMM) as classifier. The experimental results gave lower Equal Error Rate (EER) of 11.73 % with TECC features and further reduced to 10.30 % with score-level fusion of LFCC and TECC features on evaluation dataset of ASVspoof 2017 challenge version 2.0 database.

***Index Terms*—** Automatic Speaker Verification (ASV), Spoof, Replay, Reverberation, Replay Configurations (RC).

1. INTRODUCTION

The voice authentication-related applications are found to be in high demand [1]. However, the Automatic Speaker Verification (ASV) systems are exposed to different spoofing attacks [2]. The replay speech signal is one among the spoofing attacks, that requires a simple speech recording device, such as tape recorder, mobile, etc. to record the target speaker's voice from a distance [3, 4]. The replay speech signal involves several factors during recording, such as the characteristics and quality of recording device, the acoustic environment, etc. Depending on the acoustic environment, other factors are introduced in replay, such as the *reverberation*. The replay Spoof Speech Detection (SSD) task is more challenging because of the such factors and hence, these factors needs to be identified.

In this paper, we analyze the replay speech signal focusing on the reverberation. The reverberation introduces delay in the natural speech signal corresponding to different reflections that further depends on the environmental conditions. The reverberation transforms a monocomponent signal into a multicomponent one, where they are spectrally very close and hence, we cannot separate the natural components from the replay components [5]. Furthermore, in this work, we analyze natural *vs.* replay (reverberated) speech signal in time-domain and corresponding Teager energy profiles. We further analyze the individual replay configuration in terms of % EER with proposed Teager Energy Cepstral Coefficients (TECC) feature set and compare the results with state-of-the-art feature sets, such as Constant-Q Cepstral Coefficients (CQCC), Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). To the best of authors' knowledge, this is the first study of its kind reporting significance of reverberation for replay SSD task.

2. REVERBERATION IN REPLAY MECHANISM

The replay speech signal is the re-recording of the target speaker's voice captured unknowingly with the help of recording device from a distance. The recording can be done at different places, such as bedroom, balcony, canteen, office, etc. When the recording is done mainly within the closed room, the reverberation is introduced severely during replay. Reverberation is the phenomenon to resist the sound after it has been stopped as a result of multiple reflections from the surfaces, such as furniture, people, air medium, etc. within a closed surface [6]. These reflections build up with each reflection and decay gradually as they are absorbed by the surfaces of objects in the space enclosed as shown in Fig. 1(a). The reflections here are 1^{st} order (with only one deviation) and 2^{nd} order (with two deviations) from the wall, surface, etc., and direct path as shown in Fig. 1(a) without any deviations. The reflections can vary from a single deviation to many deviations.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

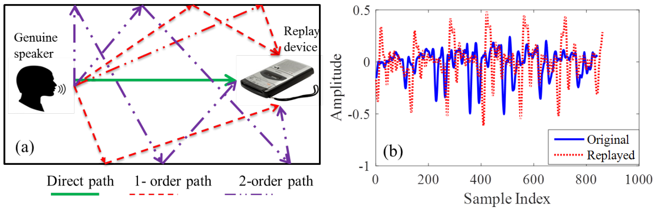

Fig. 1. (a) Schematic of reflections in room between the source and the recording (replay) device, and (b) segment of natural *vs.* replay speech signal showing the effect of reverberation.

It can be clearly observed from Fig. 1(b) that reverberation introduces delay and change in amplitude w.r.t natural speech [5]. The replay speech samples are shifted and the amplitude also varies compared to the natural speech signal. The replay signal (with reverberation) can be modeled as convolution of natural speech signal, $s(t)$, with impulse response of acoustic environment, $h(t)$ [7, 8]. The natural speech is repeated, time-shifted, and scaled for every non-zero point in the impulse response and the resulting signals are summed as shown via a schematic representation in Fig. 2. If a room do not have any signal absorbing surfaces, such as wall, roof, and floor, the signal bounce back between the surfaces and takes very long (ideally infinite) time for the signal to end. In such a room, the listener or the recording device will hear/record both the direct signal as well as the repeated reflected signal waves as shown in Fig. 1(a). If these reverberations will be more excessive, the sound will run together with a mere loss of articulation, and it becomes muddy and also garbled [6]. The time-domain speech signal are shown for both natural Fig. 3(a) and reverberated speech signals in Fig. 3(b).

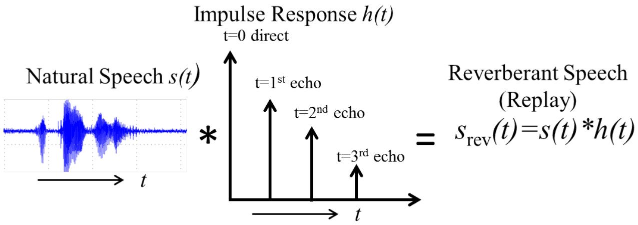

Fig. 2. Convolution of natural speech with impulse response at different echo times to obtain reverberant (replay) signal.

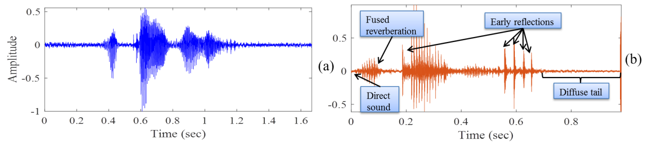

Fig. 3. Parameters involved in reverberated (replay) signal.

Discrete early reflections (1^{st} or 2^{nd} order reflections) are typically involved in the early regions of an impulse response. The reflections further become densely packed in time-domain, composing the diffuse tail (as seen in Fig. 3(b)) [9]. The time of the peak indicates how long the reflected signal will arrive at the recording device and the amplitude of the peak shows the amplitude of the reflected signal [9]. The first peak of the reverberated signal corresponds to the signal that arrives directly from the source which arrives with the shortest possible delay. The other subsequent peaks arrives because of reflections each related to its particular path that come in its way. Eventually, the reflections become sufficiently dense that they overlap in time. Because energy is absorbed by environmental surfaces with each reflection (as well as by air), longer paths produce lower amplitudes, and the overlapping echoes produce a tail in the impulse response that decays with time [9]. The impulse response is known to carry the information of the acoustic environment [10–12]. The larger rooms have few reflections resulting in slow decay of reverberated signals and the decay rates are also affected by material, such as carpet, curtains, sofa-sets, etc. Reverberation is also found to distort the structure of source signals in the spectral energy density [9, 13–15].

The Teager Energy Operator (TEO), $\Psi_d\{\cdot\}$, is used to compute the running estimate of signals energy and is given as the product of amplitude and frequency [16], i.e.,

$$\Psi_d\{y(n)\} = y^2(n) - y(n-1)y(n+1) \approx A^2\omega^2. \quad (1)$$

The Fig. 4 and Fig. 5 shows the Teager energy profiles for synthetic speech, simulated replay and natural speech. In all these cases, TEO profile shows high energy pulses around the Glottal Closure Instant (GCI), because of impulse-like excitation to vocal tract and this sudden glottal closure produces high energy and thus, TEO produces high energy around these regions [17]. Along with high Teager energy pulses, the bumps are also observed around the energy pulses, these bumps indicates the significant contribution of nonlinear effects during the speech production process [17]. The presence of bumps around the energy pulses indicates that the speech production process has the significance of nonlinear model. This is observed in the simulation experiment as shown in Fig. 4(d and h) (bumps are not observed for synthetic case). On the other hand, for speech signal the bumps around Teager energy profiles are observed as shown in Fig. 5 indicating that the natural speech production has the *nonlinear* effects.

Simulation is done to observe the effect of reverberation on

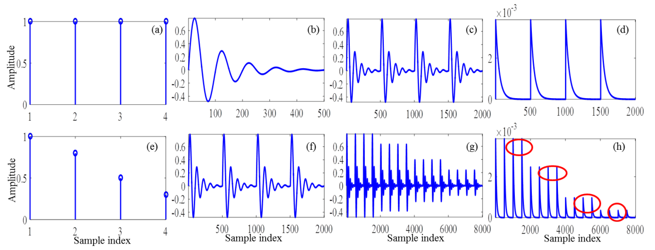

Fig. 4. (a-e) Train of impulse and echo; (b) damped sinusoid signal; (c-f) convolved signal from (a and b); (g) convolved signal from (e and f); (d-h) Teager energy profiles of (c-g).

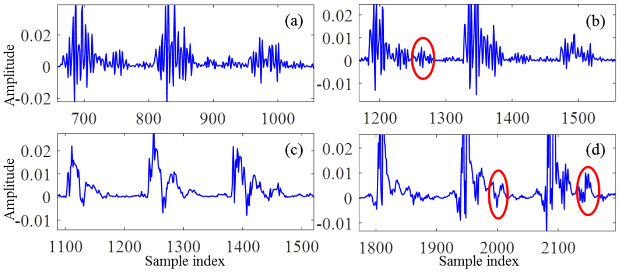

Fig. 5. Teager energy profiles of replay speech signal (a) balcony, (b) bedroom, (c) canteen, and (d) office environment (highlighted ovals shows the extra Teager energy pulse).

the Teager energy profiles in Fig. 4. The train of impulses (Fig. 4(a)) is convolved with a damped sinusoid (Fig. 4(b)) producing a convoluted signal (Fig. 4(c)). Now, to get a reverberated signal, we have convolved the convoluted signal (Fig. 4(f)) with the train of impulse echo (Fig. 4(e)) and obtained a reverberated signal with having close copies of original signal (Fig. 4(g)). The Teager energy profiles of individual signal Fig. 4(c and g) are shown in Fig. 4(d and h). We can observe that because of the reverberation (echo impulse), extra impulse-like Teager energy traces occurs between the two original Teager energy profiles (highlighted by oval). These pulses of TEO profile primarily occurs because of simulated reverberation. These extra pulses are also observed when the replay signal is recorded in a closed room, e.g., bedroom and office as observed in Fig. 5(b and d). On the other hand, energy traces are not observed for replay speech when recorded in balcony and canteen environment Fig. 5(a and c). In this Section, we studied modulations of energies estimated via TEO to emphasize the impulse that arrives because of echo/reverberation. Furthermore, we observed that for different environments, the Teager energy traces obtained are different. In particular, for a closed room (such as bedroom, office, etc.) extra energy traces are observed because of echo impulse. Hence, these observations motivated us to extract features that are based on the energy traces, and thus, proposed Teager Energy Cepstral Coefficients (TECC) discussed in the next Section.

3. FEATURE EXTRACTION PROCESS

The block diagram of TECC feature set is shown in Fig. 6. The TECC feature set earlier was used in the study of robust speech recognition where Gammatone filterbank was used [18]. The input speech signal is passed through the pre-emphasis filter as the higher frequency regions are important for replay SSD task [13]. This pre-emphasized speech signal is given to the Gabor filterbank to obtain subband filtered signals [13, 14, 19, 20]. The center frequencies are linearly-spaced in Gabor filterbank. The Gabor filterbank have *optimal* joint time-frequency resolution [21, 22]. Furthermore, these subband filtered signals are given to the TEO block to compute the running estimate of energy of each subband filtered signal. 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 along with Cepstral Mean Normalization (CMN) technique and retained first few DCT coefficients to obtain TECC feature set, followed by their Δ and $\Delta\Delta$ feature vector to obtain higher-dimensional (D) coefficients.

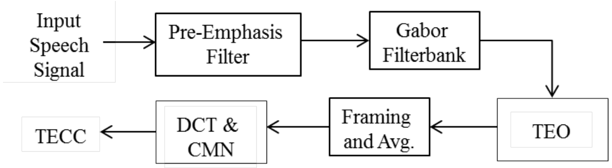

Fig. 6. Block diagram of proposed TECC feature extraction.

4. EXPERIMENTAL RESULTS

The ASVspoof 2017 challenge version 2.0 database is based on RedDots corpus and its replayed version [23–25]. The detailed statistics of the database is given in [24]. We have used GMM classifier for modeling the natural and replayed classes with 512 number of Gaussian components in GMM. The final scores are represented in terms of Log-Likelihood Ratio [13]. The details of the parameters used for feature extraction of various feature sets are as follows: CQCC features are extracted using the Cepstral Mean Variance Normalization (CMVN) technique and retained 30 static coefficients appended along with their Δ and $\Delta\Delta$ coefficients. The LFCC feature set is extracted using 40 number of subband filters in filterbank with *120*-D feature vector that includes static+Δ+$\Delta\Delta$ coefficients and for MFCC feature set, we used 40 subband filters in filterbank and extracted *39*-D feature vector. For TECC feature set, we used linearly-spaced 40 subband filters in Gabor filterbank and extracted *120*-D feature vector that includes static+Δ+$\Delta\Delta$ feature vector.

4.1. Replay Configuration (RC) with various threat levels

The level of noise in acoustic environment, playback, and recording device are assumed to be inversely proportional to the threat for ASV system pose [24]. The acoustic environment were classified into three different threat levels, namely, low, medium, and high. According to the levels of threat, the % EER of TECC feature set along with modified baseline system are shown in Fig. 7. The least % EER for all levels of threats are obtained with the proposed TECC feature set and is observed for every replay configuration.

The acoustic environment listed in [24] are the actual space in which the original speech data is replayed and rerecorded. The ASVspoof 2017 challenge version 2.0 database have 26 different environments denoted from E01-E26. Different environments have the variations with the levels of additive ambient, convolutive, and reverberation noise. The Fig. 8(a) shows the detailed % EER for all the different environmental conditions with all the feature sets on evaluation dataset. We can observe that for CQCC, MFCC, and LFCC feature sets the % EER for all the environment conditions are high compared to TECC feature set. Hence, TECC feature set (red line) shows the lower % EER for all the different environment conditions.

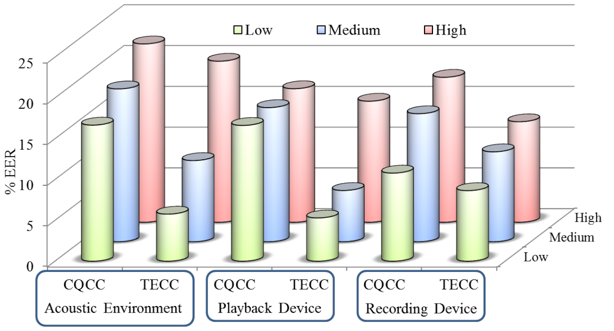

Fig. 7. Results on various replay configurations (RC) with different threat levels for CQCC and TECC feature sets.

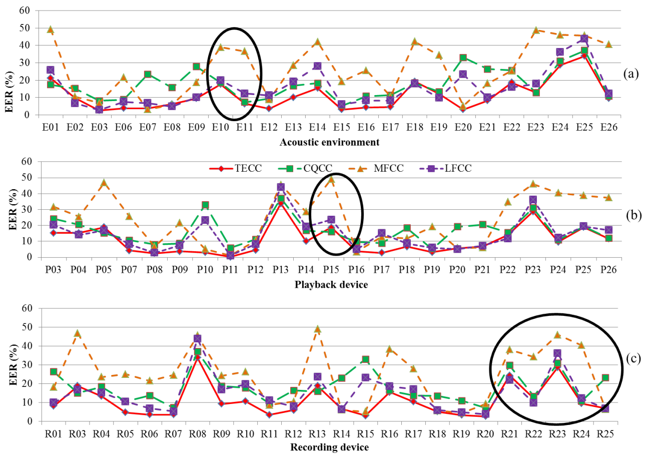

Fig. 8. Individual % EER for different acoustic environments with CQCC, MFCC, LFCC, and TECC feature sets (highlighted ovals indicates relatively better % EER by TECC for a specific acoustic environment).

Similar to different acoustic environments, there are 26 and 25 different playback and recording devices denoted by P01-P26 and R01-R25 [24]. Fig. 8(b) and Fig. 8(c) shows the detailed % EER for different playback and recording devices with all the feature sets on evaluation dataset. 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 [24]. The TECC feature set perform better in such high level threat shown by the highlighted ovals in Fig. 8. The TECC feature set shows lower % EER for all replay configurations compared to other feature sets.

4.2. Comparison of Results in % EER

We have compared the performance of TECC feature set with CQCC, MFCC, and LFCC feature sets, the results on development and evaluation sets are shown in Table 1. The CQCC feature set is the baseline system provided by the organizers of the ASVspoof 2017 challenge database [23]. We have considered LFCC feature set as other baseline, since LFCC feature set uses linear frequency scale during feature extraction process [26], and TECC feature set also uses linear frequency scale. The performance evaluation is also obtained with Detection Error Trade-off (DET) curves for all the features along with best score-level fusion given as, $S = \alpha \times S_1 + (1-\alpha) \times S_2$, where S_1 and S_2 are two feature sets and α is the fusion parameter which lies between 0 and 1 [27] of LFCC and TECC with fusion factor α=0.5 and α=0.7 on development (a) and evaluation (b) dataset as shown in Fig. 9. It is observed that the miss probability of CQCC, MFCC, and LFCC was very high for given false alarm probability which is not a good case for ASV system whereas, TECC feature set has significant decrease in miss probability.

Table 1. Comparison of results in % EER

Feature Set	Development	Evaluation
CQCC	12.81	19.04
MFCC	24.19	26.90
LFCC	16.76	13.90
TECC	9.55	11.73
TECC+CQCC	8.60	11.56
TECC+MFCC	9.55	11.73
TECC+LFCC	**8.26**	**10.30**

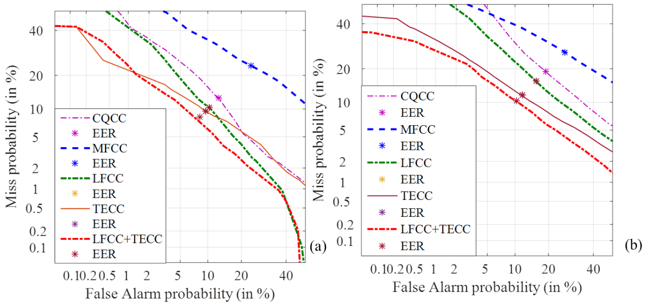

Fig. 9. Individual DET curves on development (a) and evaluation (b) dataset.

5. SUMMARY AND CONCLUSIONS

In this paper, we analyzed the effect of reverberation using TEO for replay SSD task. The delay and change of amplitude in the replay speech signal arrives because of the reverberation. Furthermore, reverberation produces the close copies of natural components making it inseparable from the replay components. The reverberated signal is also affected by the material kept in the recording environment, the shape and size of the room, the sound absorbing property of the material kept in the room, etc. Furthermore, we used the energy of subband filtered speech signal extracted from the TEO profile to calculate the % EER of individual replay configurations. The TECC feature set gave lower % EER for all replay configurations and also for different level of threats for the ASV system.

6. REFERENCES

[1] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, “Voice anti-spoofing,” in *Handbook of Biometric Antispoofing, S. Marcel, SZ Li, and M. Nixon, Eds. Springer*, 2014.

[2] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, “Spoofing and countermeasures for speaker verification: A survey,” *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] 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, 2011, Springer, pp. 274–285.

[4] FedeRerico Alegre, Artur Janicki, and Nicholas Evans, “Reassessing 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] Ixone Arroabarren, Xavier Rodet, and Alfonso Carlosena, “On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1413–1421, 2006.

[6] “Reverberation,” `https://byjus.com/physics/reverberation/`, {Last accessed: 2018-10-24}.

[7] Roman Kuc, *Introduction to Digital Signal Processing*, 1^{st} Edition, McGraw-Hill, Inc., 1988.

[8] Keisuke Kinoshita et al., “A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1–19, 2016.

[9] James Traer and Josh H. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[10] Barry Blesser and Linda-Ruth Salter, *Spaces Speak, Are You Listening?: Experiencing Aural Architecture*, MIT Press, 2009.

[11] Heinrich Kuttruff, *Room Acoustics*, 1^{st} Edition, CRC Press, 2016.

[12] Tammo Houtgast and Herman JM Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[13] Madhu R. Kamble, Hemlata Tak, and Hemant A. Patil, “Effectiveness of speech demodulation-based features for replay detection,” in *INTERSPEECH*, Hyderabad, India, 2018, pp. 641–645.

[14] Madhu R. Kamble and Hemant A. Patil, “Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection,” in *INTERSPEECH*, Hyderabad, India, 2018, pp. 646–650.

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

[16] James 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.

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

[18] Dimitrios Dimitriadis, Petros Maragos, and Alexandros 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.

[19] Petros Maragos, Thomas F. Quatieri, and James 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.

[20] Madhu R. Kamble and Hemant A. Patil, “Novel amplitude weighted frequency modulation features for replay spoof detection,” *ISCSLP*, Taipei, Taiwan, 2018.

[21] Stéphane Mallat, *A Wavelet Tour of Signal Processing. 2^{nd} Edition*, Academic press, 1999.

[22] Madhu. R. Kamble and Hemant. 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.

[23] Tomi Kinnunen et al., “The ASVspoof 2017 Challenge: Assessing the limits of replay spoofing attack detection,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1–6.

[24] Héctor Delgado, Massimiliano Todisco, et al., “ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements,” in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d’Olonne, France, 2018, pp. 296–303.

[25] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, et al., “Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, 2017, pp. 5395–5399.

[26] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma, “Linear versus mel frequency cepstral coefficients for speaker recognition,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, 2011, pp. 559–564.

[27] Zahid Akhtar, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli, “Evaluation of multimodal biometric score fusion rules under spoof attacks,” in *IEEE International Conference on Biometrics (ICB)*, New Delhi, India, 2012, pp. 402–407.