Subband Channel Selection using TEO for Replay Spoof Detection in Voice Assistants

Harsh Kotta, Ankur T. Patil, Rajul Acharya, and Hemant A. Patil
Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology
(DA-IICT), Gandhinagar-382007, Gujarat, India.
E-mail: $\{harsh\_kotta, ankur\_patil, rajul\_acharya, hemant\_patil\}@daiict.ac.in$

***Abstract*—Recently, there is an increase in the demand for Voice Assistants (VAs) due to their convenience in accessing and controlling the household devices. To make VAs user-friendly, less strict speaker verification constraints are imposed onto them which makes VAs highly vulnerable to spoofing attacks. In this paper, authors propose the design of front-end countermeasure system against replay spoofing attack for VAs that make use of microphone array to capture spatial diversity. We exploit this microphone array information by proposing a novel approach of the subband channel selection using mathematical structure of Teager Energy Operator (TEO). These selected subband channels are used to compute proposed Teager Energy Cepstral Coefficients (TECC$_{max}$) feature set. With this approach, we gain significant improvement in the performance of replay attack detection task on VAs against the baseline feature set, i.e., Constant-Q Cepstral Coefficient (CQCC). Results indicate an absolute reduction in Equal Error Rate (EER) of 4.11% and 8.66% on development and evaluation set, respectively, of ReMASC dataset. Authors also performed classifier-level fusion of GMM, and LCNN-based back end classifiers using proposed TECC$_{max}$ feature set and obtained absolute reduction of 5.98% and 10.67% on development and evaluation sets, respectively.**

Keywords: Teager Energy Operator, TECC, ReMASC.

I. Introduction

Voice Assistants (VAs) are becoming ubiquitous due to their ease of operation and controllability of personal and household appliances. However, their convenience makes them highly prone to spoofing attacks. Variety of spoofing attacks are reported in the literature which can cause great threat to VAs [1], [2], [3], [4], [5], [6]. This issue raises a need to design robust anti-spoofing technique. In this study, authors propose a countermeasure system against the replay attacks on VAs by exploiting the Teager energy-based features. The experiments are performed on the recently released *Realistic Replay Attack Microphone Array Speech Corpus* (ReMASC) corpus [7]. This corpus is specifically designed for developing countermeasure system against the replay spoofing attack on VAs using microphone array.

Earlier, several datasets were released to address the spoofing attacks on Automatic Speaker Verification (ASV) systems along with countermeasure systems for respective datasets [8], [9], [10]. However, the design of the countermeasure systems for VAs is quite different than that used for the ASV systems. Such differences are illustrated in [7]. Their simplicity in ASV had made VAs vulnerable to spoofing attacks. In replay spoof, distortions due to recording, playback devices, and environments are incorporated along with genuine speech sample. Based on this modeling of the speech signal, several countermeasure systems are designed [11], [12].

In this work, subband filtering and Teager Energy Operator (TEO) for feature representation, where TEO is energy estimating operator. The feature sets derived from TEO have been successful in many speech applications because of its capability to capture the nonlinearity in the speech signal, high temporal resolution, and noise suppression capability [13], [14], [15]. Due to these desirable properties, TEO has been used to develop Teager Energy Cepstral Coefficients (TECC) for speech recognition application [16]. TEO can further be utilized to estimate the amplitude and frequency modulations (AM-FM) in the speech signal [17], [18], [19], [20], [21]. This time resolution property can be used to track rapid changes in signal's energy within a glottal cycle [15]. TECC is found to be most successful feature set to develop whisper speech recognition system [22]. Many studies reported efficacy of TEO-based features for SSD task [23], [24], [25], [26], [27], [28].

TEO has the capability to detect the noise components present in the signal hence, it is used here for selecting the most distorted subband channel signal, obtained from the microphone array which can be used effectively for replay SSD task. In this paper, we propose to use TECC due to it's inherent capability of capturing the acoustic reverberation as discussed in our recent work [29]. Our key idea involves the extraction of TECC features based on the subband channel selection. Here, *a channel* refers to a speech signal obtained from a single microphone in given microphone array. Furthermore, a signal filtered through the bandpass filter in the filterbank is referred here as *subband signal*. Availability of microphone array enables selection of appropriate subband channel to maximize the reverberation cues via spatial diversity for the SSD task. This difference once identified can act as an effective countermeasure against replay spoofing attack on VAs as in ReMASC dataset.

II. TECC using Subband channel selection

A. Replay Noise Analysis using TEO

In [15], a non-linear differential operator was used to track energy analogously to a mass-spring system. This energy operator is known as TEO. Let the discrete-time signal $y(n)$ be expressed as $y(n) = A\cos(\omega n + \theta)$, then the running estimate

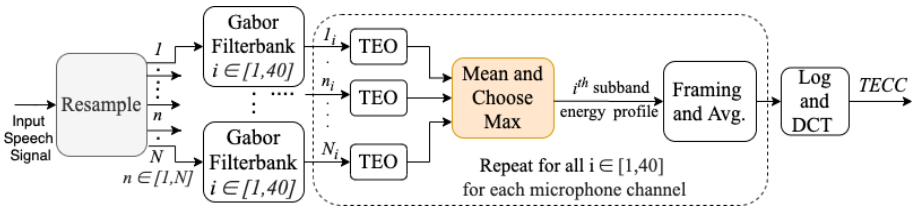

Fig. 1. Block diagram of the proposed TECC feature set extraction for replay SSD on VAs.

of energy of the signal using TEO $\psi\{.\}$ can be written as:

$$\psi[y(n)] = y^2(n) - y(n-1)\cdot y(n+1) = A^2\sin^2(\omega) \approx A^2\omega^2. \quad (1)$$

In eq. (1), the approximation used is $sin(\omega) \approx \omega$ for small values of ω. Energy Separation Algorithm (ESA) was developed in order to find the individual contribution by amplitude and frequency components of the signal. Speech signal is produced because of pulsatile vortical flow interactions with the boundaries of the vocal tract system [17]. Speech resonances depend upon local vocal tract cavities that emphasizes certain frequencies and de-emphasizes the others [19]. As speech signal itself is composed of several multicomponent resonances and hence, in order to apply ESA, there is a need to separate resonances by bandpass filtering using an appropriate filterbank. Therefore, TEO profile of each of these subband signals is calculated using eq. (1). Let us assume that $x(n)$ is composed of original clean speech signal, $s(n)$, and noise component, $r(n)$, i.e., $x(n) = s(n) + r(n)$. Then,

$$\psi[x(n)] = [s(n)+r(n)]^2 - [s(n+1)+r(n+1)]\cdot [s(n-1)+r(n-1)]. \quad (2)$$

By definition of average cross-TEO [17], [30]:

$$\psi_{cr}^{avg}(s(n),r(n)) = s(n)r(n) - 0.5[s(n+1)r(n-1) + r(n+1)s(n-1)]. \quad (3)$$

Using eq. (2) and eq. (3), we obtain,

$$\psi[x(n)] = \psi[s(n)] + \psi[r(n)] + 2\psi_{cr}^{avg}(s(n),r(n)). \quad (4)$$

Let additive noise, $r(n)$, be the zero-mean wide-sense stationary (WSS) Gaussian random process. Then, applying expectation operator $\mathbb{E}\{\cdot\}$, we get,

$$\mathbb{E}\{\psi[x(n)]\} = \mathbb{E}\{\psi[s(n)]\} + 2\mathbb{E}\{\psi_{cr}^{avg}(s(n),r(n))\}, \quad (5)$$

since $\mathbb{E}\{\psi[r(n)]\} = 0$ as $r(n)$ is a zero-mean WSS process. According to the analysis given in [30], eq. (5) can be written as:

$$\mathbb{E}\{\psi[x(n)]\} = \mathbb{E}\{\psi[s(n)]\} + error. \quad (6)$$

In eq. (6), the first term in RHS will be constant for each subband filtered signal, the only varying term will be the $error$ term which is expected to be responsible for acoustic noise in replay speech. In SSD task, the $error$ term can be maximized by selecting the subband channel, which consists of maximum error, i.e., acoustic noise. Hence, we are selecting the subband channel having maximum average TEO profile (i.e., LHS of eq.(5)). This is the key novelty in our proposed work.

B. TECC Feature Extraction Scheme for SSD Task in VAs

The details of feature extraction scheme is depicted in Fig. 1. The input speech signal is recorded via different microphone arrays, each having different sampling frequencies which is due to the use of different recording devices. Hence, each utterance is first resampled to 16 *kHz*. Due to microphone array, speech signal representation consist of N channels, where N is the number of microphones in microphone array. The resulting multichannel input is then passed through a linearly-spaced Gabor filterbank which has optimal time-frequency resolution, i.e., $\sigma_t^2\sigma_w^2 = 0.5$, where σ_t^2 and σ_w^2 are variances or uncertainties in time and frequency domain, respectively [31]. We have used 40 subband filters ($i \in [1, 40]$) for this purpose. Thus, there will be i subband filtered signals obtained from each channel. The TEO is then applied on each subband filtered signal to estimate the subband energy for each channel. Average subband energy for each channel is estimated to select the subband signal with maximum average. This is done so as to maximize the acoustic noise in the selected subband of the n^{th} channel. This is followed by framing and averaging operation. For extracting TECC feature set, we have used a window length of 20ms, and hopping size of 10ms. This is followed by logarithmic operation to compress the dynamic range. Finally, Discrete Cosine Transform (DCT) is applied to obtain static TECC feature set. Velocity and acceleration features are appended so as to encapsulate transitional information. The proposed feature set is abbreviated as TECC$_{max}$ as we chosen the subband channel having maximum energy. This novel approch for subband channel selection shows the improvement in the performance of the countermeasure system over TECC extracted from the single channel. The performance of this feature sets is discussed in Section IV.

III. Experimental Setup

A. Dataset

In this work, ReMASC dataset is used which aims to develop effective countermeasures against replay spoofing attack on VAs [7]. The details of the dataset collection strategies can be found in [7]. From the available dataset, $\sim$ 25500 utterances are used. We partitioned the dataset into 3 subsets, i.e., training, development, and evaluation set. Data distribution

is shown in Table I. This publicly available dataset consists of recordings from 44 subjects. The data partition in Table I consists of 22, 17, and 20 speakers in training, development, and evaluation subset respectively. Most of the speakers in training and development subset are overlapping. However, speakers selected in evaluation set are disjoint to that of training and development subset. The recordings have been performed in four different environments. The same proportion of all recording environments is maintained in the all the subsets.

TABLE I
DESIGN OF REMASC DATABASE. AFTER [7]

	Training	Development	Evaluation
Genuine	2820	924	3308
Spoof	7392	1884	9203
Total	10212	2808	12511

B. Feature Sets and Classifiers

We used two classifiers in this study, namely, Gaussian Mixture Model (GMM), and Light Convolutional Neural Network (LCNN). The state-of-the-art features for the SSD task, namely, Constant-Q Cepstral Coefficients (CQCC), and Linear Frequency Cepstral Coefficients (LFCC) has been used along with Mel Frequency Cepstral Coefficients (MFCC) for comparison using GMM classifier [32], [33]. The performance of our proposed feature set, i.e., TECC$_{max}$ is compared against these mentioned feature sets. The CQCC, LFCC, MFCC, and TECC are extracted with 90-*D*, 60-*D*, 42-*D*, and 120-*D*, feature vectors, respectively. All the feature sets consists of static coefficients appended with velocity and acceleration coefficients. In TECC feature set, we have chosen the subband channel having maximum energy in order to capture maximum distortion due to acoustic (replay) noise so that it becomes highly discriminative feature for SSD task. In order to validate our key idea, we also performed experiments by choosing the subband channels having minimum energy. We referred this feature set as TECC$_{min}$. Furthermore, experiments are also performed by extracting the TECC from the single channel of the microphone array. This experiment can be considered as random selection of the subband channel. In Table II, this channel selection scheme is abbreviated as TECC$_{random}$.

GMM is trained with 512 Gaussian mixtures along with training set with various feature sets. Log-Likelihood scores (LLk) are obtained by providing the test utterances as input to trained *defense* models [34]. The Equal Error Rate (EER) is chosen as the performance measure to evaluate the countermeasure systems [35]. We employed LCNN as an alternate classifier or defense model. This deep neural network architecture uses Max-Feature-Map activation [36], [37]. This architecture was also used for the replay SSD task in [38]. To train this defense model, CQCC and TECC$_{max}$ feature sets are used. In ReMASC dataset, duration of the utterances are of varying size. To train the LCNN network, we require consistent feature representation. To achieve this, we replicate the audio samples if their duration is less than 4 seconds. Otherwise, they are truncated to 4 seconds. We deploy the defense model with a learning rate of 10^{-3}, and a batch size of 32 samples with ADAM optimizer.

We also performed the classifier-level fusion for the likelihood scores obtained from different systems to investigate the possible complementary information contents of different classifiers. Likelihood scores obtained from two different systems are fused as:

$$LLk_{fused} = \gamma \cdot LLk_{system1} + (1-\gamma) \cdot LLk_{system2}, \quad (7)$$

where $LLk_{system1}$, and $LLk_{system2}$ are the log-likelihood scores obtained by system-1 and system-2, respectively. The fusion parameter $\gamma \in [0, 1]$ decides the relative importance of the two systems.

IV. EXPERIMENTAL RESULTS

A. Spectrographic Analysis of Genuine vs. Replay Spoof

We analyzed the Constant-Q Transform gram (CQT-gram) against the TEO-gram, to observe the possible discriminative capability of these two feature sets. As shown in Fig. 2, Panel-I and Panel-II shows the TEO-gram, and CQT-gram, respectively. Fig. 2(a) and Fig. 2(c) shows the TEO-gram of genuine and spoof speech signals, respectively. Whereas, Fig. 2(b) and Fig. 2(d) shows the CQT-grams for genuine and spoof speech signals, respectively. The frequency scale in TEO-gram is linear, whereas it is non-linear in CQT-gram. From Fig. 2, considering a typical range of 4KHz to 8KHz, it can be observed that there is spectral smearing (or blurring) in the higher frequency regions for both CQCC and TECC feature sets between genuine signal and its spoof counterpart. However, it is more clearly (better resolution) observed in TEO-gram than the CQT-gram which is shown by highlighted regions in Fig. 2. It can be also observed that CQCC feature set has high frequency resolution in lower frequency region as compared to the high frequency region. This discrimination, however, is not observed in case of TECC feature set as both low and higher frequency regions are equally highlighted. This property can act as an effective discriminative feature for differentiating genuine and spoof utterances on VAs as in a replay attack AM-FM modulations can occur both in low as well as in high frequency regions.

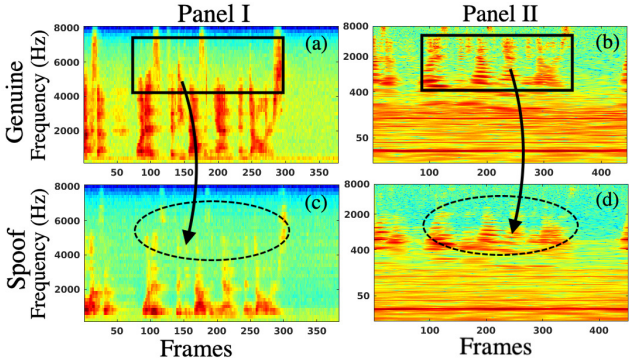

Fig. 2. Spectrogram plot of TECC (Panel I) *vs.* CQCC (Panel II) feature sets : (a),(b) for genuine speech signal, and (c),(d) for spoofed speech signal.

B. Results using Individual vs. Fused Systems

To evaluate the efficacy of our approach of subband channel selection by Teager energy tracking, we performed experiments using seven individual systems as shown in Table II. It can be observed that the TECC$_{max}$ feature set performs significantly better than the other cepstral features using GMM classifier. In particular, the absolute reduction in EER of 4.11% and 8.66% is obtained for TECC$_{max}$-GMM system over CQCC-GMM system, i.e., the relative improvement in EER of 37.15% of proposed approach over the baseline system, on evaluation set. To validate effectiveness of selecting maximum energy subband channel, we performed experiment on selecting minimum energy subband channel to design TECC$_{min}$. Furthermore, the performance of the system developed using TECC$_{random}$ is obtained in between two extremes, i.e., in between TECC$_{min}$ and TECC$_{max}$. The considerable difference in the performance is observed between TECC$_{max}$ and TECC$_{min}$ frontend. To validate the efficacy of our approach, we used another classifier, i.e., LCNN along with baseline feature set CQCC and proposed TECC$_{max}$. LCNN shows better performance with TECC$_{max}$ than the CQCC feature set. Furthermore, classifier-level fusion of the GMM and LCNN systems for proposed TECC$_{max}$ feature set reduces an EER to 14.59% and 12.64% for development and evaluation set, respectively. Thus, it achieves 45.77% relative improvement over the baseline system for the evaluation set. It can also be concluded that energy-based TECC feature (TECC$_{min}$, TECC$_{max}$, or TECC$_{random}$) set encapsulates more effective information for replay SSD task as compared to the other feature sets shown in Table II.

TABLE II
RESULTS (IN % EER) ON REMASC DATASET

System	Dev	Eval
CQCC-GMM *(Baseline)*	20.57	23.31
LFCC-GMM	28.89	26.31
MFCC-GMM	36.43	31.53
TECC$_{min}$-GMM	19.89	16.78
TECC$_{max}$-GMM (A)	**16.46**	**14.65**
TECC$_{random}$-GMM	18.57	16.61
CQCC-LCNN	22.31	25.88
TECC$_{max}$-LCNN (B)	17.98	16.84
A + B	**14.59**	**12.64**

C. Results on Environment-Dependent and Independent Scenarios

Experiments are also performed for environment-dependent, and independent scenario. For environment-dependent case, target environment is already seen by the defense model. In this case, each environment is partitioned into two disjoint and speaker-independent sets of roughly the same size. The results obtained using CQCC and TECC$_{max}$ feature sets with GMM classifier, are reported in Table III. Particularly for this scenario, we reported the results with application of the Cepstral Mean Variance Normalization (CMVN) to each feature utterance as it has shown significant improvement in the performance for both the feature sets. This needs further investigation and is an open problem. It can be observed that the TECC$_{max}$ performs better than the CQCC for all environments.

In environment-independent scenario, defense model is trained on any of the three environments and tested on the fourth environment. Results in Table III shows that the proposed TECC$_{max}$ feature set performs better for three unseen environments, namely, environment-A, B, and C, whereas both the feature sets shows the poor performance on environment-D indicating environment-D could not be expressed as linear combination of the other environments.

TABLE III
RESULTS (IN % EER) FOR ENVIRONMENT-DEPENDENT *vs.* INDEPENDENT CASE ON REMASC DATASET ON GMM CLASSIFIER

	Feature Set	Env-A	Env-B	Env-C	Env-D
Env-Dependent	CQCC	23.27	42.62	12.96	15.85
	TECC$_{max}$	**13.39**	**27.92**	**10.30**	**9.06**
Env-Independent	CQCC	35.65	40.89	35.95	49.99
	TECC$_{max}$	**26.76**	**35.43**	**31.79**	**49.98**

Env = Environment

D. Results using Detection Error Trade-off (DET) Curves

Analysis is also performed by observing the DET plots shown in Fig. 3 [39]. It can be observed that the TECC$_{max}$ feature set is consistently performing well. For development set, all the DET curves for TECC$_{max}$ are inclined towards the lower indices of miss probability. Whereas, this trend is observed for TECC$_{max}$-LCNN system on the evaluation set. Having lesser miss probability is desirable attribute for a good SSD system.

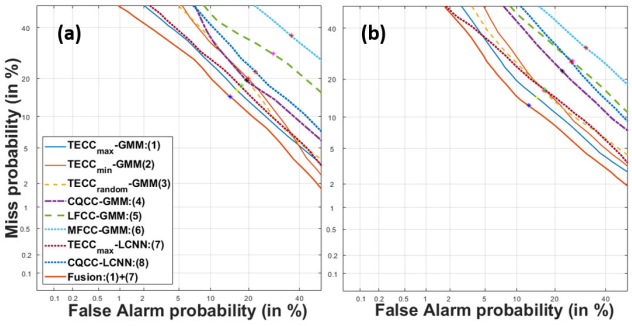

Fig. 3. DET curves for the performance of the systems shown in Table III (a) development set, and (b) evaluation set. Legends of Fig. 3 (b) are similar as in Fig. 3 (a).

V. SUMMARY AND CONCLUSIONS

In this study, we proposed a novel strategy of selecting the subband channel based on the maximum noise energy estimated, via TEO of the signal. Along with genuine speech signal, the replayed spoof speech signal consists of additional components, i.e., impulse responses of the recording and replay environments and devices. The effect of these additional components are characterized by acoustic noise in the channel. To make this characteristic of spoof speech signal more distinct, we have chosen maximum energy subband channel in the feature representation. Using this strategy of subband channel selection having maximum energy, a significant improvement

is obtained in the performance over the baseline CQCC-GMM system. It also performed well in environment-dependent scenario. For environment-independent cases, TECC$_{max}$ performs better than the baseline, however, results are not promising for the deployment of the system in practice. This issue can be addressed in near future.

REFERENCES

[1] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in 25^{th} *USENIX Security Symposium*, Austin, USA, August 2016, pp. 513–530.

[2] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in 4^{th} *ACM Workshop on SPSM*, Scottsdale, USA, Nov. 2014, pp. 63–74.

[3] Y. Gong and C. Poellabauer, "An overview of vulnerabilities of voice controlled systems," 1^{st} *International workshop on Security and privacy for Internet-of-Things, Orlando, United States*, April 2018.

[4] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Security and Privacy Workshops (SPW)*, San Francisco, USA, May 2018, pp. 1–7.

[5] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," in *9th USENIX Workshop on Offensive Technologies (WOOT-15)*, Washington, DC, USA, 2015.

[6] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *arXiv preprint arXiv:1711.03280*, 2017.

[7] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," in *INTERSPEECH*, Graz, Austria, 2019, pp. 2355–2359.

[8] Z. Wu, T. Kinnunen, N. 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.

[9] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "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. 1–8.

[10] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 1008–1012.

[11] P. A. Tapkir, A. T. Patil, N. Shah, and H. A. Patil, "Novel spectral root cepstral features for replay spoof detection," in *APSIPA-ASC*, Honolulu, Hawaii, USA, 2018, pp. 1945–1950.

[12] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, September 2014, pp. 1–6.

[13] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

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

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

[16] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 3013–3016.

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

[18] 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.

[19] P. Maragos, J. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[20] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3245–3265, 1993.

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

[22] D. T. Grozdic and S. T. Jovicic, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2313–2322, 2017.

[23] A. T. Patil, R. Acharya, P. A. Sai, and H. A. Patil, "Energy separation-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection," *Proc. INTERSPEECH 2019*, pp. 2898–2902, 2019.

[24] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection." in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 12–16.

[25] M. R. Kamble and H. A. Patil, "Effectiveness of Mel scale-based ESA-IFCC features for classification of natural *vs*. spoofed speech," in *International Conference on Pattern Recognition and Machine Intelligence*. Kolkata, India: Springer, December 2017, pp. 308–316.

[26] 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.

[27] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection." in *Interspeech*, 2018, pp. 641–645.

[28] M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 106–110.

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

[30] S. Lefkimmiatis, P. Maragos, and A. Katsamanis, "Multisensor multiband cross-energy tracking for feature extraction and recognition," in *ICASSP*, Las Vegas, USA, April 2008, pp. 4741–4744.

[31] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

[32] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[33] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[34] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[35] W. Millar, J. Oglesby, M. Pawlewski, and J. Tang, "The assessment of speaker verification systems," *Proc. of Inst. of Acoustics*, vol. 14, pp. p423–p423, 1992.

[36] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[37] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *International Conference on Machine Learning, Atlanta, USA*, June 2013.

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

[39] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 1895–1898.