Energy Separation Based Features for Replay Spoof Detection for Voice Assistant

Gauri P. Prajapati, Madhu R. Kamble and Hemant A. Patil
Speech Research Lab
Dhirunhai Ambani Institute of Information and Communication Technology (DA-IICT)
Gandhinagar, Gujarat, India
(gauri_prajapati, madhu_kamble, hemant_patil)@daiict.ac.in

***Abstract*—Voice Assistant (VA) now-a-days plays a very important role for the smart home applications. However, the VA along with ease also brings security issue too, such as possibility of being attacked by replay, hidden voice commands, etc. This paper presents replay Spoof Speech Detection (SSD) system for VA using Energy Separation Algorithm (ESA)-based features to capture Instantaneous Amplitude and Frequency Cepstral Coefficients (i.e., ESA-IACC and ESA-IFCC), and Gaussian Mixture Model (GMM) as a pattern classifier. Teager Energy Operator (TEO) has the characteristics to suppress the noise and hence, it is robust to noise sensitivity. For noisy acoustic environment, the ESA-based features that employ TEO perform well compared to the clean environment. We performed the experiments on the ReMASC database, which contains four different acoustic environments. Proposed features performed better in clean and noisy environments. In addition, to obtain possible complementary information, we performed score-level fusion of ESA-IACC and ESA-IFCC that resulted in low Equal Error Rate (EER) for different environments. Furthermore, we compared our proposed feature sets with Constant-Q Cepstral Coefficients (CQCC), and Linear Frequency Cepstral Coefficients (LFCC) resulting in an relative improvement of approximately 21.88 % for clean environments and 66.34 % for noisy environments (in EER), respectively.**

***Index Terms*—Replay Spoofing, Voice Assistant (VA), Teager Energy Operator (TEO), Energy Separation Algorithm (ESA).**

I. Introduction

Biometric authentication, such as fingerprint scanning, retinal scanning, face detection, and voice recognition, etc. are used in intelligent personal assistants. Internet of Things (IoT) system used speakers voice as the user-machine interaction. Voice Assistant (VA), such as Google Assistant, Amazon Alexa, SIRI, etc. are widely used inside our homes to control the appliances with the help of smart speakers [1], [2]. Eventhough there are plenty of use of voice assistants (VA), they raise several security concerns, such as spoofing attacks. Spoofing is defined as the imposter speaker impersonate as a genuine speaker to get access of a secured system or data [3]–[5]. There are different spoofing attacks, such as speech synthesis (SS) [6], voice conversion (VC) [7], [8], replay [9], [10], impersonation [11], and twins [12] primarily in the context of Automatic Speaker Verification (ASV).

Few attacks that dominates voice assistant are hidden voice commands [13], self-triggered attacks [14], and audio adversarial examples [2]. The study reported in [15]–[17] states that any attack can tend to harsh/serious losses, e.g. an attacker may do online shopping, a burglar might enter a house by replaying the owners voice, make financial transactions on behalf of the owner without direct authorization, and so on.

Few studies reported on countermeasures to alleviate spoofing attacks on VA [2], [18]. The attacks on the VA also depends on implementation techniques pf VA, namely, Basic Voice Replay attack, Operating System-Level attack (e.g., Ally attack [19], Google Voice Search (GVS) attack [17]), Hardware-level attack (i.e., Dolphin attack [20], Intentional Electromagnetic Interference (IEMI) attack [21]), Machine learning-level attack [14], etc. These techniques are mainly based on the replay signal and hence, it is required to increase the performance of the system. To handle these attacks, one can use the approach used to develop counter measures for spoofing attacks on ASV systems. However, the methods which works excellent for ASV systems may or may not perform well for VA. One major difference is that the ASV systems assume that the user is in controlled environment whereas in case of VA, there are several different acoustic environments where the user gives command to the VA. The ASV systems use more strict configurations of ASV model. In the contrast, for making them easy to use, VAs may not have the strict ASV models. Few VAs do not have by default such as, Xiaomi MI AI a smart home control system. This makes it easy for the imposter to do a replay attack to give malicious commands to the VAs to access the system. As a result, VAs need more robust defense model to distinguish between natural *vs.* replayed speech.

In this paper, we are exploring our earlier proposed feature sets, i.e., Energy Separation Algorithm (ESA)-based Instantaneous Amplitude and Instantaneous Frequency (IA-IF) component of a signal for VA [22]–[24]. The experiments are performed on a recently available Realistic Replay Attack Corpus for Voice Controlled Systems (ReMASC) corpus, which is more competent in replay detection for VA. In particular, we studied the effect of different acoustic environments along with the IA-IF-based features on VA. The Teager Energy Operator (TEO) has the capability to suppress the noise and hence, in the signal degradation conditions, the performance of the TEO-based features is generally better compared to the other features.

II. Energy Separation Algorithm (ESA)-Based Features

The Energy Separation Algorithm-Instantaneous Amplitude Cepstral Coefficients (ESA-IACC), and Energy Separation Algorithm-Instantaneous Frequency Cepstral Coefficients (ESA-IFCC) feature extraction process is shown in Fig. 1. The input speech signal is passed through the pre-emphasis filter to enhance the higher frequency regions. Furthermore, the speech signal is passed through the linearly-spaced Gabor filterbank in order to obtain narrowband filtered signals in order to estimate the instantaneous Teager energy profile along with ESA [25]–[28]. The ESA provides the corresponding Instantaneous Amplitude and Instantaneous Frequency (IA-IF) components of a narrowband filtered signals [29]. These estimated IA and IF components are further processed to obtain corresponding speech segments with a window length of 25 ms along with a shift of 10 ms followed by logarithm operation to compress the data. To obtain a low-dimensional representation that has compact energy, Discrete Cosine Transform (DCT) is applied along with Cepstral Mean Variance Normalization (CMVN) to reduce the channel mismatch/distortion conditions [30]. Finally, retained few DCT coefficients, i.e., ESA-IACC , and ESA-IFCC appended along with their Δ and $\Delta\Delta$ features to obtain higher-dimensional feature vector. Please note that the experiments performed in Section IV for ESA-IFCC feature set are extracted without applying pre-and-post-processing, as this process gave better results than the other feature extraction parameters used.

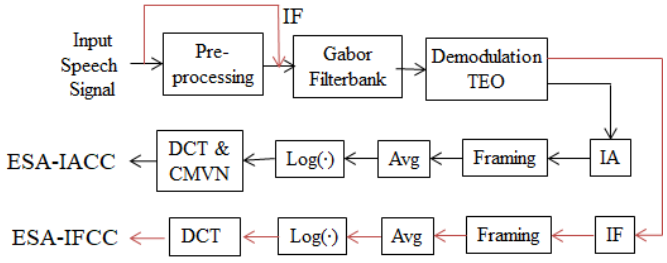

Fig. 1. Block diagram of ESA-IACC, and ESA-IFCC feature sets. After [23].

In addition, we also observed and compared the spectral energy densities of traditional Short-Time Fourier Transform (STFT) spectrogram with the spectral energy obtained from Teager energy-based approach as shown in Fig. 2. The comparison is shown for all the acoustic environments from the ReMASC database, in particular, outdoor (Panel I-II), indoor 1 (Panel III-IV), indoor 2 (Panel V-VI), and vehicle (Panel VII-VIII). For outdoor environment, it can be observed from the Fig. 2 that the spectral energy is not preserved for both Panel I and Panel II. In addition, for high frequency regions, we observe high energy. However, with Teager energy-based approach, we preserve the mid and higher frequency information compared to the traditional spectrogram. The spoof signal of corresponding outdoor environment shows much more distortion in spectral energies compared to its natural counterpart. As the recording is done in the open outdoor area, it is indeed possible that the signal carries different types of noise along with it and hence, the performance degrades for outdoor environment. Similarly, we observe the spectral energy differences for other acoustic environment. In particular, for indoor 2 and vehicle acoustic environments, it can be clearly observed that the spectral energy obtained from the Teager energy-based approach preserves much more information about the formants and harmonics compared to the traditional spectrogram and hence, the performance for these environment is better compared to the other environments (discussed in Section IV).

III. Experimental Setup

We performed the expperiments on the publicly available ReMASC corpus, which is specifically developed for the VAs that includes near-field and far-field speech. The brief discussion of acoustic environment is mentioned next. The dataset consists of two disjoint sets, namely, Core and Quick evaluation set [31]. The statistics of the ReMASC database is given in Table I.

TABLE I
Statistics of ReMASC Corpus. After [31]

Environment	Core (Training) set			Quick Eval. (Test) set
	Replay utterances	Genuine utterances	Total utterances	Utterances
Env A	1259	707	1966	257
Env B	5659	1069	6728	720
Env C	2040	835	2875	318
Env D	3858	2657	6515	1062
Total	**12816**	**5268**	**18084**	**2357**

A. Acoustic Environments:

The database was created in four different acoustic environments, namely [31]:

- *Outdoor Environment (Env A):* The data is collected in a student plaza with various background noises to imitate the noisy outdoor environment conditions.
- *Indoor Environment 1 (Env B):* Data is recorded from a quiet study room using three different placement positions of recorders: Center of the room, against the wall, and corner of a room. This environment allows us to emulate the situation of flexible device, and speaker positions.
- *Indoor Environment 2 (Env C):* In reality, while using VAs, there are several background sounds that are mixed with the original voice commands. To imitate this, the data is collected at a lounge with TVs, and music players running in background.
- *Vehicle Environment (Env D):* The data is collected in a moving vehicle, once in a silent parking area (with the engine is off), and once when car is moving. This will allow us to analyze the effects of vehicle-based VAs.

The ReMASC corpus was developed for three different tasks, namely, *Task 1*: Mismatch training and testing condition, *Task 2*: Environment-independent, and *Task 3*: Environment-dependent with different speakers in training and testing set [31]. In this study, we are focusing on the Task 3, i.e., environment-dependent (which will have the same acoustic

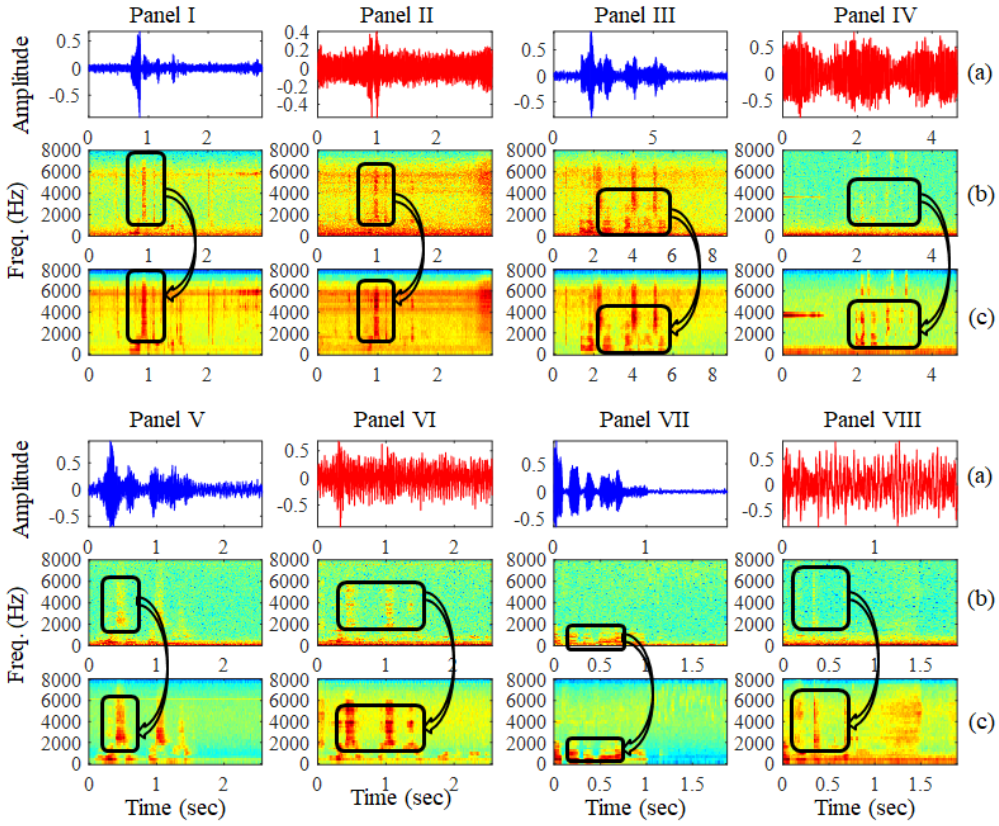

Fig. 2. Spectral energy densities obtained from the traditional STFT, and Teager energy-based approach for different acoustic environments. Panel I-II: Outdoor, Panel III-IV: Indoor 1, Panel V-VI: Indoor 2, Panel VII-VIII: Vehicle. (a) Time-domain signal in different acoustic environment along with their corresponding, (b) traditional STFT, and (c) Teager energy-based approach.

environment during training and testing, however, the number of speakers taken in training will be different from that of speakers selected during testing).

B. Baseline Systems:

The organizers of the ReMASC corpus provided the baseline system with CQCC feature set as front-end features along with Gaussian Mixture Models (GMM) as the classifier at back-end. The CQCC feature set was extracted with *30*-static coefficients appending with *30*-Δ and *30*-$\Delta\Delta$ resulting in total *90*-D feature vector [32], [33]. In addition, we used LFCC feature set to compare with the proposed feature sets. The LFCC feature set was extracted using *60*-D feature vector that includes (*20*-static+Δ+$\Delta\Delta$) [34]. The ESA-IACC, and ESA-IFCC feature sets are extracted with $f_{max} = 8000$ Hz, and $f_{min} = 10$ Hz with linear frequency scale in the Gabor filterbank. The ESA-IFCC feature set is extracted without applying pre-emphasis and CMVN techniques. On the other hand, the ESA-IACC feature set is extracted using pre-emphasis and CMVN techniques. From our earlier studies, we found that the ESA-IACC and ESA-IFCC feature sets perform better with the 40 number of subband filtered signals with *120*-D feature vector (that includes static+Δ+$\Delta\Delta$). If we reduce the number of subband filters or feature dimension, the performance of the Spoof Speech Detection (SSD) for VA degrades.

GMM is used to map each class as a weighted sum of *I* multivariate Gaussians [35]. It is given by $p(x|\lambda)$=$\sum_{k=1}^{N} w_k p_k(x)$, where w_k is the k^{th} mixture weight, and $p_k(x)$ is a *D*-variate Gaussian density function with mean vector μ_i, and covariance matrix, $\sum_i$. The model parameter is defined by λ. The likelihood scores are found for natural *vs.* replayed speech. The decision against hypothesis is based on Log-Likelihood Ratio (LLR)=$\log \frac{P(X|H_o)}{P(X|H_1)}$, where $P(X|H_o)$ and $P(X|H_1)$ are the likelihood scores of natural and replay speech, respectively. The score-level fusion is performed to combine possible complementary information and is given by:

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

where LLK$_{feature1}$, and LLK$_{feature2}$ is log-likelihood score of feature1 and feature2, respectively, and α is the fusion parameter or weight ($0 < \alpha < 1$).

IV. Experimental Results

This Section describes the experiments performed on the Task 3, i.e., environment-dependent. We observed the Power Spectral Density (PSD) obtained after applying TEO on the small speech segment for the natural, and its corresponding spoof speech signal as shown in Fig. 3. The PSD for different environment, namely, (a) outdoor, (b) indoor 1, (c) indoor 2, and (d) vehicle shows the difference from its natural counterpart. In particular, we observe differences in the PSD plots for the indoor 2 and vehicle environments, which indeed help us to detect the spoof signal (which is also observed from

our experimental results discussed in the next sub-section). Furthermore, the performance for ESA-IACC, and ESA-IFCC feature sets is enhanced in the next sub-section.

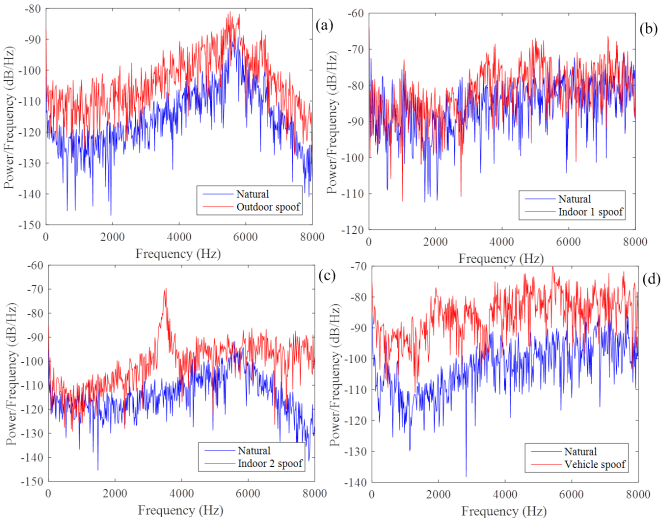

Fig. 3. Power Spectral Density (PSD) of natural (blue color) speech segment, and its corresponding replay (red color) speech recorded in (a) outdoor, (b) indoor 1, (c) indoor 2, and (d) vehicle acoustic environment.

A. Results on ESA-IACC Feature Set

The experiments are performed by varying the number of subband filters in a Gabor filterbank from 40 to 100 for all the acoustic environments (as shown in Fig. 4). It can be observed from Fig. 4 that the effect of number of subband filters indeed degrades the performance for a particular environment, and at the same time, it performs better for the other environments. In particular, for Env B (indoor 1), it can be observed that we get high EER for all the number of subband filters. The possible reason behind it could be the environmental conditions, which the replay spoof speech is recorded. As indoor 1 environment is quiet study room and hence, the replay spoof signal will be similar to the natural speech resulting in less discrimination in replay speech from its natural counterpart, degradation in the SSD performance. On the other hand, the spoof signal when recorded in other environments, i.e., indoor 2 and vehicle are able to detect as these environments are having noise added in the replay signals that is used as the discrimination feature from the natural speech because the Teager energy-based features have noise suppression capability, and its features are robust to noise sensitivity.

B. Results on ESA-IFCC Feature Set

Similar to experiments in Section IV-A, we performed the experiments for ESA-IFCC feature set with varying the number of subband filters from 40 to 100. It can be observed from Fig. 5 that for acoustic environment indoor 2 and vehicle, the EERs are low compared to the other two acoustic environments, namely, outdoor and indoor 1. With increasing the number of subband filters in a filterbank, the EER decreases and hence, it proves that the narrowband filtering for extracting TEO-based features are essential for detecting spoof speech signals, which also further depends on the acoustic environment (i.e., noisy *vs.* clean environment).

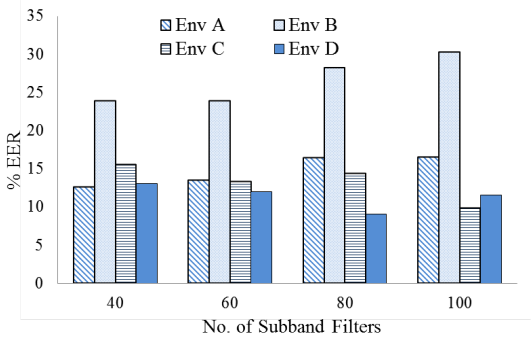

Fig. 4. Results in EER (%) for ESA-IACC features sets with varying the number of subband filters for different acoustic environments.

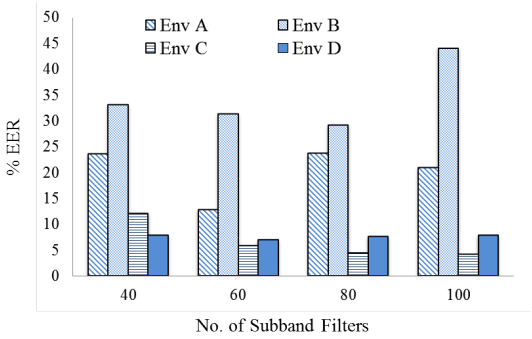

Fig. 5. Results in EER (%) for ESA-IFCC features sets with varying the number of subband filters for different acoustic environments.

C. Results with Score-Level Fusion

We further compared results for our proposed feature sets with the baseline system along with LFCC feature sets in Table II. It can be observed that the ESA-IACC and ESA-IFCC feature set performed better for Env A, Env C, and Env D. However, for Env B, the ESA-based feature sets fail to detect the replay speech signal.The reason for this is discussed in the section IV containing the PSD discussion. When compared to CQCC and LFCC feature sets, we obtained much lower EERs for all the environments apart from indoor 1. Furthermore, we performed the score-level fusion of ESA-IACC and ESA-IFCC feature sets to further improve the performance of the replay SSD task. The score-level fusion indeed helped to get the lower EER than individual EERs for both feature sets. For outdoor, indoor 2, and vehicle environment, the score-level fusion gave EER of 11.92 %, 2.07 %, and 5.18 %. It represents that the score-level fusion of both the feature sets capture complementary information that helped us to improve the replay SSD performance than the individual feature sets.

TABLE II
COMPARISON (IN % EER) WITH OTHER FEATURE SETS ALONG WITH SCORE-LEVEL FUSION RESULTS (IN % EER)

Feature sets	EER			
	Env A	**Env B**	**Env C**	**Env D**
CQCC	15.26	**17.41**	6.15	6.59
LFCC	22.44	24.41	15.97	18.24
ESA-IFCC	19.36	29.11	4.06	6.22
ESA-IACC	12.59	23.84	9.81	9.11
ESA-IFCC+ESA-IACC	**11.92**	21.00	**2.07**	**5.18**

V. Summary and Conclusions

In this paper, we studied the importance of different acoustic environments for Voice Assistants (VAs). In particular, we found that the noisy and clean environment indeed affect the performance to detect the replay speech signal from its natural counterpart. We used Energy Separation Algorithm (ESA)-based Instantaneous Amplitude and Instantaneous Frequency feature sets to detect the replay signals. The speech signal when recorded in noisy environment has distortions, however, using the ESA-IFCC feature sets; this type of replay signals are classified from it's natural counterpart. On the other hand, when the signals are recorded in the clean environment, they are difficult to detect as they might be similar to the natural signal and hence, very less differences in them are observed. Thus, for the clean environment, our proposed feature sets fails to classify the replay signal and hence, more detailed analysis and study is required to detect the replay signal in such scenarios, which forms our immediate future work.

References

[1] M. B. Hoy, "Alexa, SIRI, cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[2] L. Blue, L. Vargas, and P. Traynor, "Hello, is it me you're looking for? differentiating between human and electronic speakers for voice interface security," in *Proceedings of the* 11^{th} *ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pp. 123–133, 2018.

[3] 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 (SLTC) Newsletter*, pp. 2013–05, 2013.

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

[5] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, pp. 1–8, 2020.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

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

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

[9] 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.

[10] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, 2016.

[11] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Proc.*, (Hong Kong), pp. 145–148, 2004.

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

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

[14] 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 ({USENIX} Security 16)*, pp. 513–530, 2016.

[15] E. Alepis and C. Patsakis, "Monkey says, monkey does: Security and privacy on voice assistants," *IEEE Access*, vol. 5, pp. 17841–17851, 2017.

[16] X. Lei, G.-H. Tu, A. X. Liu, K. Ali, C.-Y. Li, and T. Xie, "The insecurity of home digital voice assistants–amazon alexa as a case study," *arXiv preprint arXiv:1712.03327, {Last Accessed: 02, March 2020}*, 2017.

[17] 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 *Proceedings of the* 4^{th} *ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, (New York, United States), pp. 63–74, 2014.

[18] Y. Gong and C. Poellabauer, "Protecting voice controlled systems using sound source identification based on acoustic cues," in *IEEE* 27^{th} *International Conference on Computer Communication and Networks (ICCCN)*, (Hangzhou, China), pp. 1–9, 2018.

[19] Y. Jang, C. Song, S. P. Chung, T. Wang, and W. Lee, "A11y attacks: Exploiting accessibility in operating systems," in *ACM SIGSAC Conference on Computer and Communications Security*, (Scottsdale, Arizona, USA), pp. 103–115, 2014.

[20] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, (Dallas, TX, USA), pp. 103–117, 2017.

[21] C. Kasmi and J. L. Esteves, "Iemi threats for information security: Remote command injection on modern smartphones," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 6, pp. 1752–1755, 2015.

[22] 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)*, pp. 106–110, IEEE, 2017.

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

[24] M. R. Kamble and H. A. Patil, "Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection," in *INTERSPEECH*, (Hyderabad, India), pp. 646–650, 2018.

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

[26] P. Maragos, J. F. Kaiser, T. F. Quatieri, *et al.*, "On separating amplitude from frequency modulations using energy operators," in *Proc. ICASSP*, vol. 2, pp. 1–4, 1992.

[27] K. Vijayan, V. Kumar, and K. S. R. Murty, "Feature extraction from analytic phase of speech signals for speaker verification," in *INTERSPEECH*, (Singapore), pp. 1658–1662, 2014.

[28] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Comm.*, vol. 81, pp. 54–71, 2016.

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

[30] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *IEEE ICASSP*, (Hong Kong, China), pp. I–656–659–I, 2003.

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

[32] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, vol. 25, (Bilbao, Spain), pp. 249–252, 2016.

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

[34] X. Zhou *et al.*, "Linear *vs.* mel frequency cepstral coefficients for speaker recognition," in *IEEE ASRU*, (Hawaii, USA), pp. 559–564, 2011.

[35] 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.