Teager Energy Cepstral Coefficients for Audio Deepfake Detection

Ritik Mahyavanshi*, C.V. Mahesh Reddy†, Arth J. Shah*, and Hemant A. Patil*
* Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India
* E-mail : {202001030, 202101154, hemant_patil}@daiict.ac.in
† Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana.
† E-mail : cvnreddym9@gmail.com

***Abstract*—Audio deepfakes have emerged as a significant concern, as these convincingly mimicked synthetic audio and can be exploited to manipulate targeted groups by altering speech information. No prior research has explored the use of Teager Energy Operator (TEO)-based features for ADD task. This paper discusses a novel technique for distinguishing deepfake audio from real ones using Teager Energy Cepstral Coefficients (TECC) features, which captures the energy fluctuations of audio signals, making it a promising approach for detecting real *vs.* deepfake audio. For classification, the ResNet-50 model is employed in combination with the FoR (Fake or Real) dataset. Experimental results demonstrate the efficacy of this method, achieving an Equal Error Rate (EER) of approximately 7.53 % and an accuracy of 92.55 % on static TECC feature sets. The findings indicate that TECC features outperform traditional techniques, such as Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). This superior performance highlights the significant potential of TECC features for further research in deepfake detection.**

***Index Terms*- Audio Deepfake Detection, Teager Energy Cepstral Coefficients, FoR Dataset.**

I. Introduction

Deepfake is a type of deep learning technique to create or manipulate video and audio content, typically to depict someone saying or doing something that never actually happened. One real-life example of deepfake technology impacting individuals involves a case, where in 2019, criminals used Artificial Intelligence (AI)-generated audio to impersonate a CEO's voice and issue fraudulent instructions to transfer funds. It highlighted the importance of better security measures and awareness to prevent the misuse of deepfake technology in sensitive situations. The rise of deepfake technology has extended beyond visual media, posing significant challenges to the authenticity of audio recordings. Audio deepfakes, generated through sophisticated deep learning models, can convincingly mimic human speech, leading to potential misuse in various domains [1]. This study focuses on an innovative approach for Audio Deepfake Detection (ADD). It leverages Teager Energy Cepstral Coefficients (TECC) derived from Teager Energy Operator (TEO) in combination with a ResNet-50 classifier. The approach is evaluated using the standard Fake or Real (FoR) dataset [2].

Researchers previously used various techniques for ADD task. Many studies have applied different models to identify fake audio clips. In [3], they suggested using the Wav2vec2 model to train a siamese neural network for safeguarding anti-spoofing models from attacks.

The focus in ADD has been on identifying suitable audio features, broadly divided into are used and learning-based features. Features, such as Mel Frequency Cepstral Coefficients (MFCC) [4], [5] and Linear Frequency Cepstral Coefficients (LFCC) [6], [7] closely mimic human auditory traits and concentrate on low frequency information crucial for speech detection tasks. The emergence of learning-based audio features in recent years has garnered significant attention for ADD research. Utilizing whisper audio features [8] has shown promise in detecting synthetic speech, surpassing handcrafted features with the vast amount of supporting data available for the whisper system. Self-supervised learning-based audio features have also displayed benefits in ADD tasks by leveraging pre-trained data from diverse sources to help differentiate between genuine and fake speech effectively even in complex situations and thus, performing well on varied datasets.

In our work, we have proposed to use TECC features as in previously used methodology, such as MFCC and LFCC features, which captures various details in audio signals through different filter window lengths across frequency bands but as it is primarily used for music analysis and in comparison; these features which have good match with human auditory characteristics are more suitable for ADD systems have a only drawback of more sensitivity towards noise. In our approach of using TECC features having more resistant to noise as TECC, which is derived by TEO naturally having noise suppression capability making TECC a potential feature set for ADD systems.

In our approach, we utilize TECC features, which are inherently resistant to noise due to the mathematical structure of TEO from which they are derived. This noise resistance makes TECC a promising area for research in ADD systems. We use the extracted TECC features as inputs for the ResNet-50 classifier [9], which is adapted for audio classification tasks. ResNet-50's deep architecture, renowned for its success in image classification, is modified to process TECC features, enabling it to discern between genuine *vs.* fake audio recordings. The FoR dataset, comprising a diverse collection of real and deepfake audio samples, is used to train and validate the proposed detection system. Experimental results demonstrate that the combination of TECC features with the

ResNet-50 classifier achieves high accuracy in identifying audio deepfakes, surpassing traditional ADD methods.

The the rest of the paper is organised as follows: In Section II gives explanation about the TEO and TECC. Section III gives description of experimental setup used in this study, and Section IV presents experimental results of our research followed by summary and conclusions in the last Section V.

II. Proposed Methodology

TEO is employed to capture non-linear aspects of energy variations in audio signals, providing a means to extract proposed TECC features. These featuers are effective in identifying fluctuations in the running estimate of energy patterns of speech, which are indicative of nonlinearities in production of natural speech.

A. *TEO*

TEO is a nonlinear operator utilized in signal processing to estimate the instantaneous energy of a signal. Its effectiveness lies in its ability to detect changes in both amplitude and frequency, making it particularly useful for the non-linearity, which define the naturality of a speech production mechanism.

Mathematically, TEO for a discrete-time signal $x[n]$ is defined as [10]:

$$\psi(x[n]) = x^2[n] - x[n-1] \cdot x[n+1]. \tag{1}$$

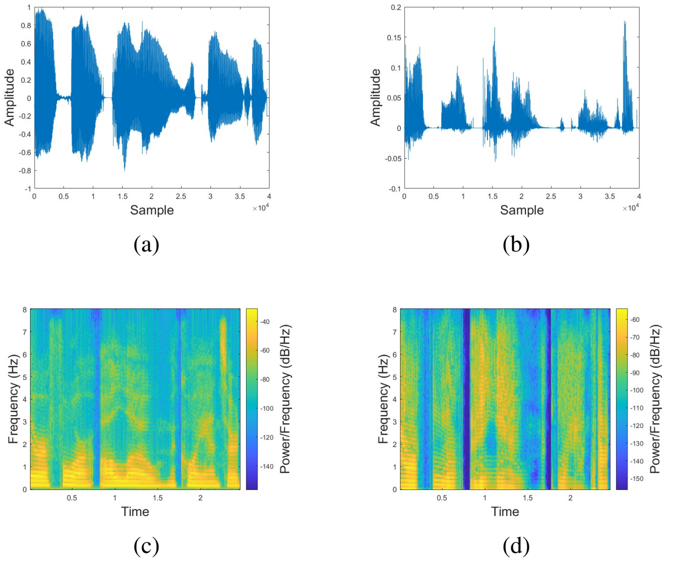
Here, $\psi(x[n])$ represents the transformed signal via TEO containing the instantaneous energy information, $x(n)$ being the discrete-time signal, and n represents the sample index.

The TEO can estimate the amplitude envelope of amplitude modulated (AM) signals and the instantaneous frequency of frequency modulated (FM) signals. It can also be used for signals with changing amplitude and frequency (AM-FM signals). The TEO profile of a audio signal is shown in Fig. 1.

B. *TECC Feature Extraction*

The functional block diagram of TECC extraction is show in the Fig. 2. Here, the input speech signal is passed through a pre-emphasis highpass filter having a Z domain system function, $H(z) = 1 - \alpha z^{-1}$; with a typical value of α (i.e. α is zero in H(z)) is set to 0.97. It is used in speech processing to enhance the quality and intelligibility of the speech signal. It works by amplifying the high frequency components of the speech signal, which typically have lower energy compared to the low frequency components. This emphasis on high frequencies helps to improve the signal-to-noise ratio (SNR), making the speech signal clearer, especially in noisy environments.

Now, the signal is passed through the Gabor filterbank [11], which consists of two fuctions: a Gaussian function and a sinusoidal waveform. The Gaussian function, $\exp(-b^2t^2)$, forms a bell-shaped envelope around the input signal, creating a window for the sinusoidal waveform, $\cos(\omega_c t)$, to be modulated. This windowing effect allows a specific range of frequencies to pass through, effectively subband filtering the signal.



(a)

(b)

(c)

(d)

Fig. 1: (a) Time-domain waveform, (b) corresponding TEO profile of the input signal shown in Fig 1 (a), (c) spectrogram of the signal shown in Fig. 1 (a), and (d) TEO spectrogram of the Fig. 1 (b).

The factor b in the Gaussian function determines the spread of the bell-shaped envelope, impacting how much of the input signal is encompassed by the window. A larger b results in a wider spread, capturing more of the input signal, while a smaller b results in a narrower spread, capturing less of the input signal. The term $\exp(-b^2t^2)$ represents the Gaussian function, while $\cos(\omega_c t)$ represents the sinusoidal waveform. The factor b in the Gaussian function affects the bell-shaped envelope's spread over the input signal. The impulse response of the Gabor filter $h(t)$ is given by:

$$h(t) = \exp(-b^2t^2)\cos(\omega_c t).$$

In the context of deepfake audio detection, the Gabor filterbank can be used to extract specific features from the audio signal. By adjusting the parameters of the Gaussian function and the sinusoidal waveform, it is possible to isolate and analyze different frequency components of the signal, aiding in the identification of synthetic or manipulated audio.

Now, TEO is applied to the formed signal in order get the TEO profile of the input signal. Here, all the frequency fluctuations and the instantaneous energy of the input signal can be observed. TEO profile of both real and fake audio is shown in the Fig. 1.

The framing and averaging step is crucial for breaking down the continuous speech signal into smaller parts called speech framer. This process involves splitting the audio into short, overlapping frames. Each frame is looked at separately under the assumption of short-term stationarity (and assumption of linear time-invariant (LTI) system), making it easier to spot subtle inconsistencies or anomalies typical of deepfake audio and efficienty exploit linear signal prediction algorithms.

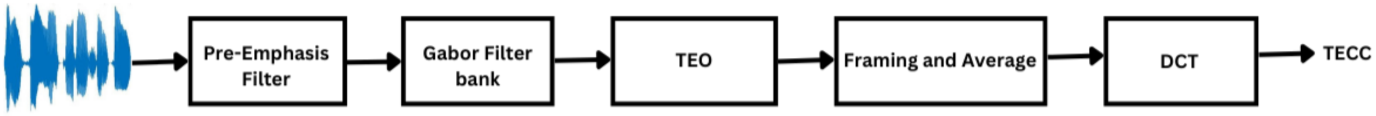


Fig. 2: Fuctional Block Diagram of TECC Feature Extraction After. [10].

Averaging within these frames helps in standardizing the signal, reducing noise, and emphasizing important features that may show tampering. By studying these frames, we can catch abnormal variations and artifacts introduced by deepfake techniques that often struggle to maintain natural coherence over brief time periods.

The Discrete Cosine Transform (DCT) phase is crucial for enhancing feature extraction in ADD. The DCT is applied to the log filter-bank subband energies, converting them into the cepstral-domain. This transformation captures essential frequency-domain features while reducing redundancy. The resulting TECC are then utilized to identify abnormalities and artifacts unique to deepfake audio, as these coefficients are highly sensitive to subtle energy variations introduced by synthetic speech methods. The application of DCT ensures that the extracted features are decorrelated, enhancing their effectiveness in distinguishing between genuine and fake audio samples. In particular, DCT is applied for feature vector decorrelation, dimensionality reduction of underlying TECC features vector, and energy compaction. Ultimately, the TECC feature encapsulates the TEO profile of the input signal, providing a robust representation for detecting audio deepfakes.

In real audio, glottal closures [12] cause distinct and sharp spikes in the TEO profile, reflecting the natural rhythm of the speaker's vocal fold vibrations. The energy released during each closure significantly contributes to the clarity and naturalness of the speech. Conversely, fake audio often simulates these spikes with less pronounced or irregular energy-levels during glottal closures. Synthetic methods struggle to replicate the natural dynamics of glottal closures, resulting in less distinct energy spikes or piles in the TEO profile. Within the regions between closures, real audio exhibits smoother and more gradual energy fluctuations, reflecting the controlled damping characteristics of human speech. This stability in energy patterns over time underscores the consistent nature of genuine vocal fold vibrations and glottal airflow. In contrast, fake audio shows more erratic energy fluctuations within these regions, with rapid rises and falls that *disrupt* natural continuity. This inconsistency in energy transitions highlights the artificial nature of the generated speech and its inability to sustain energy-levels as effectively as real speech.

The differences in energy fluctuations and persistence between real *vs.* fake audio provide crucial indicators for detecting audio deepfakes. Analyzing TEO profiles allows for the identification of genuine speech characteristics, such as stable and persistent energy patterns in real audio, contrasting sharply with the erratic and rapidly decaying energy in fake audio. These distinctions are instrumental in developing effective deepfake detection algorithms, ensuring the integrity and authenticity of audio content.

III. Experimental setup

A. Dataset Used

Several datasets featuring synthetic speech have been previously published, yet there are compelling reasons for introducing the dataset outlined in this study. Existing datasets predominantly lack utterances generated from the latest deep-learning-based speech synthesis algorithms. Furthermore, their volume of utterances is often insufficient for training complex neural network models effectively. Moreover, most focus on spoofed utterance detection for speaker verification systems. The FoR dataset [2], tailored for ADD, is a meticulously curated collection of audio recordings aimed at advancing research in identifying synthetic or manipulated speech. It encompasses a diverse array of audio samples, including authentic human speech and various types of deepfake audio produced through techniques, such as Text-to-Speech (TTS) and voice conversion (VC).

TABLE I: Distribution of FoR Corpus. After [2].

Subset	**Fake**	**Real**
Training	26,924	26,938
Testing	2,370	2,264
Validation	5,398	5,399

FoR Dataset , which contains more than 87,000 synthetic utterances as well as more than 111,000 real utterances. Such a dataset is fundamental for research in synthetic speech detection, since it contains enough data to train the most complex deep learning algorithms, which makes it more suited for Deep Neural Network (DNN), such as ResNet-50 classifier.

B. Pattern Classifier Used

We performed experiments using ResNet-50 as a two-class classifier, where the two classes are real and fake speech signals [9]. ResNet-50, a variant of the Residual Networks (ResNet) architecture, excels in various image recognition tasks due to its depth and sophisticated design. The architecture consists of four main blocks, each progressively deeper and more complex. The first block has three convolutional layers that capture low-level features, such as *edges* and *textures*. Each convolutional layer is followed by batch normalization and a ReLU activation function to stabilize the learning process and introduce non-linearity.

The second block contains four convolutional layers, building on the features detected in the first block and identifying more complex patterns. The third block, with six convolutional layers, captures detailed and abstract features essential for distinguishing subtle differences in the data. The fourth and final block has three convolutional layers, consolidating the features learned in the previous blocks.

After these four blocks, the feature maps undergo global average pooling, reducing their spatial dimensions by computing the average of each feature map and summarizing the spatial information. This reduction helps prevent overfitting and reduces computational complexity. The architecture then includes a fully-connected layer that integrates the pooled features into a single vector, which is passed through a softmax activation function to output a probability distribution over the different classes, providing the final classification results.

ResNet-50's design, with its depth and residual connections that facilitate gradient flow, makes it highly effective for deep feature extraction and robust learning. This architecture has been successfully applied in numerous applications, including deepfake detection, where it benefits from its ability to learn and distinguish intricate patterns in visual data.

IV. Experimental Results

We focused on the TECC features and explored a thorough dimensional analysis comparison with existing approaches to evaluate its efficacy and efficiency. The dimensional analysis revealed that TECC offers superior performance in key metrics, such as accuracy, precision, and computational complexity. Additionally, a detailed latency period analysis was performed to assess the real-time applicability of TECC.

A. Effect of Dimension of TECC Feature Vector

To optimize audio classification models, we have thoroughly investigated both static and dynamic conditions for the TECC features. This exploration is crucial for refining model performance, as selecting the ideal conditions for feature extraction can significantly impact accuracy. Our analysis revealed that among various configurations, the static TECC features with dimension 30 provides the most effective results, achieving a remarkable accuracy rate of 92.55% and EER of 7.53%. Given its superior performance, we have considered this static configuration as the optimal condition for further classification purposes. The dynamic features show an accuracy rate of 91.45% with 28 dimension. While notable, this does not surpass the highest accuracy achieved under static conditions. Thus, only static conditions are preferred for classification.

B. Comparison with Cepstral Features

Analysis of different feature sets under these optimal conditions reveals that TECC leads with an accuracy of 92.55%, an F1-Score of 92.52%, and an EER of 7.53%. In comparison, MFCC achieves 66.61% accuracy, 63.05% F1-Score, and 32.63% EER, while LFCC records 49.87% accuracy, 35.33% F1, and 49.01% EER. This demonstrates that TECC outperforms both MFCC and LFCC, as shown in Table II.

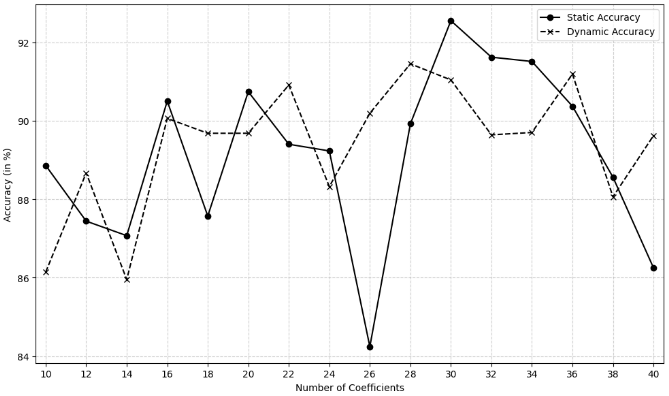


Fig. 3: Effect of Dimension of TECC Feature Vector.

TABLE II: Performance Comparison with Existing Features

Feature Set	Accuracy	F1-Score	EER
LFCC	49.87%	35.33%	49.01%
MFCC	66.61%	63.05%	32.63%
TECC	**92.55%**	**92.52%**	**7.53%**

C. Analysis of Latency Period

To analyze the proposed TECC feature vector and optimize speech frame length, we need to reduce storage requirements. Latency analysis helps to determine the minimum data needed to effectively train the model, thereby optimizing storage requirements. We explored different latency conditions for various features, including TECC, MFCC, and LFCC. The graph shown in Fig. 4 labels the accuracy performance of TECC, LFCC, and MFCC w.r.t. varying speech frame lengths.

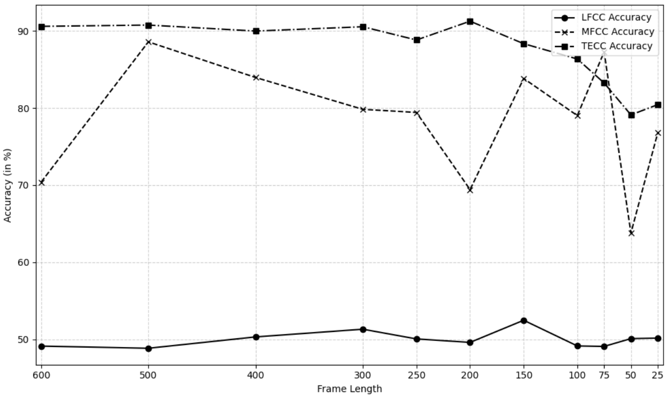


Fig. 4: Analysis of Latency Period for LFCC, MFCC, and TECC.

TECC maintains an impressive accuracy range between 79.11% and 91.26%. This high level of performance is especially notable at a frame length of 200, where TECC achieves its highest accuracy of 91.26%. In contrast, MFCC, although capable of high accuracy, demonstrates unreliable performance with substantial drops, and LFCC consistently underperforms. The results demonstrated that TECC significantly reduces

TABLE III: Comparison With Related Works in the Literature

Source	Feature Set	Classifier	Accuracy (in %)
[13]	MFCC-20	XGB	59
		KNN	62
		RF	62
		SVM	67
[2]	Spectral analysis (Brightness, Hardness, Density)	NB	67.27
		DT	70.26
		RF	71.47
		SVM	73.46
[2]	STET, MFCC, and CQT	VGG19	89.79
Proposed	**TECC**	**ResNet-50**	**92.55**

the latency period compared to existing methods, ensuring faster and more reliable outcomes. These findings highlight TECC's capability to improve performance and reduce delays, establishing it as a reliable solution for real-world applications. Therefore, TECC's ability to deliver robust and dependable accuracy across different frame lengths makes it the superior feature extraction technique for ensuring consistent and reliable results.

D. Comparison with Existing Studies

In this study, we assessed the performance of various classifiers and feature sets for a classification task. The results, as summarized in Table III, reveal a notable range in effectiveness across different combinations. Among the feature sets evaluated, the TECC feature set used with the ResNet-50 classifier achieved the highest accuracy of 92.55 %. This superior performance underscores the efficacy of the TECC feature set in conjunction with the ResNet-50 classifier compared to other combinations. The findings highlight the significant impact that the choice of feature set and classifier can have on classification accuracy, with the proposed TECC feature set demonstrating the best overall performance in our experiment.

V. Summary and Conclusions

The study underscores the superior performance of combining TECC features with a ResNet-50 classifier for detecting audio deepfakes. TECC's ability to capture nonlinear characteristics enhances the model's accuracy in differentiating between real *vs.* fake audio. Evaluation on the standard FoR dataset shows that TECC consistently achieves the lowest EER and highest accuracy compared to existing MFCC and LFCC features. Specifically, TECC's lower EER values indicate its superior performance in minimizing both false acceptance and false rejection rates, and its higher accuracy underscores its effectiveness in correct classification. This suggests that TECC, when used with a ResNet-50 classifier, not only reduces classification errors more effectively than MFCC and LFCC but also offers robust performance in real-world audio deepfake detection. Future research could build on this success by integrating additional features and classifiers to further enhance detection capabilities and address more advanced deepfake techniques.

Appendix: noise superrasion of TEO

The TEO has demonstrated its noise suppression capability in previous analyses, particularly for speech recognition tasks, including handling various noises and person recognition in noisy settings. Here, we examine TEO's effectiveness in suppressing additive noise. Let $s[n]$ represent the clean speech signal and $\hat{s}[n] = s[n] + w[n]$ denote the noisy speech signal, where $w[n]$ is a zero-mean additive noise component. We present the TEO profiles for both $s[n]$ and $w[n]$:

$$\begin{aligned} \Psi\{s[n]\} &= s^2[n] - s[n-1]s[n+1], \\ \Psi\{w[n]\} &= w^2[n] - w[n-1]w[n+1]. \end{aligned} \tag{2}$$

The TEO profile for the noisy speech signal $\hat{s}[n]$ is calculated as:

$$\begin{aligned} \Psi\{\hat{s}[n]\} &= \hat{s}^2[n] - \hat{s}[n-1]\hat{s}[n+1] \\ &= (s[n] + w[n])^2, \\ &\quad - (s[n-1] + w[n-1])(s[n+1] + w[n+1]) \\ &= s^2[n] + 2s[n]w[n] + w^2[n], \\ &\quad - s[n-1]s[n+1] - s[n-1]w[n+1], \\ &\quad - w[n-1]s[n+1] - w[n-1]w[n+1]. \end{aligned} \tag{3}$$

Moreover from [10], we know that:

$$E[\Psi\{\hat{s}[n]\}] \approx E[\Psi\{s[n]\}]. \tag{4}$$

The Eq. (4) indicates that when TEO is applied to a noisy signal with zero-mean additive noise, it can effectively suppress the noise. Thus, TEO exhibits a notable noise suppression capability.

Acknowledgements

This study has been funded by Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the sponsorship of project 'BHASHINI', (Grant ID: 11(1)2022-HCC.(TDIL)). Authors thank authorities of of DA-IICT Gandhinagar for their kind support and cooperation to carry out this research work.

References

[1] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, 2022, ISSN: 1999-4893. [Online]. Available: https://www.mdpi.com/1999-4893/15/5/155.

[2] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.

[3] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," in *Proc. INTERSPEECH*, 2023, Dublin, Ireland, pp. 2808–2812.

[4] A. Hamza, A. R. R. Javed, F. Iqbal, *et al.*, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022, 9996362.

[5] A. Qais, A. Rastogi, A. Saxena, A. Rana, and D. Sinha, "Deepfake audio detection with neural networks using audio features," in *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, 2022, Hyderabad, India, pp. 1–6.

[6] P. Kawa, M. Plata, and P. Syga, "Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection," *arXiv preprint arXiv:2206.13979*, 2022, $\{\text{Last Accessed Date : } 22^{nd} July, 2024\}$.

[7] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for add 2022," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 9226–9230.

[8] Y. Yang, H. Qin, H. Zhou, *et al.*, "A robust audio deepfake detection system via multi-view feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, Seoul, Korea, pp. 13 131–13 135.

[9] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio deepfake approaches," *IEEE Access*, vol. 11, pp. 132 652–132 682, 2023.

[10] 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)*, Singapore, 2019, Brighton, UK, pp. 2607–2611.

[11] R. Hammouche, A. Attia, S. Akhrouf, and Z. Akhtar, "Gabor filter bank with deep autoencoder based face recognition system," *Expert Systems with Applications, pp. 11*, vol. 197, 2022.

[12] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2011.

[13] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering, vol. 47, pp. 3447–3458*, 2021.