Significance of Lower Frequency Regions for Audio Deepfake Detection

Arth J. Shah, and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, India
E-mail: {202101154, hemant_patil}@daiict.ac.in

***Abstract*—Deepfake audios are created using deep learning methods. Audio security, due to deepfake audio creation, has been a severe issue in recent days. Many techniques have been explored to detect attacks by fake audio generators. Similar to Spoofed Speech Detection (SSD), audio deepfake systems also rely on characteristics of live speech to detect whether the audio is deepfake or real, however, they aim to fool humans instead of voice biometrics system. As the attackers are free to mount various types of audio attacks, Audio Deepfake Detection (ADD) plays a vital role in defending against fake AI/deep learning-generated audio attacks. For this study, we explored various lower frequency-based acoustic features, such as Generalized Morse Wavelet (GMW)-based features in combination with Mel spectrogram-based features, using Convolutional Neural Network (CNN)-based classifier for ADD task. For performance comparison, we employed the traditional spectrogram. Further, feature-level data fusion (instead of score-level) of proposed GMW-based with Mel spectrogram feature gave 1.93% improvement in overall accuracy.**

***Index Terms* – Audio Deepfake, Morse Wavelet, Lower Frequency Resolution, Mel Spectrogram, Speech Intelligibility.**

I. Introduction

Deepfake is a content, where the original recording of audio, video, and photo has been replaced by computer-generated fake signal using deep learning algorithm. Deepfake was first enlightened by a user named "Deepfakes" on Reddit in 2017 [1]. Due to advances in technology, the activities of criminals have increased several folds w.r.t. deepfake attacks. With the increase in the availability of facilities, deepfakes have been created more critically, which makes them harder to detect. In recent years, numerous such attacks based on deepfake have been made [2], [3], which drew the attention of researchers. For such an advanced generated deepfake, we need a strong detection system that can classify audio more accurately. The existing baseline approaches used acoustic features for the Audio Deepfake Detection (ADD) task, whose success motivated us to explore other acoustic features for the ADD task. Moving through the characteristics of audio, many acoustic features have been successfully used for various speech security-based applications. Many times, speech signals are analyzed using spectrogram of speech signals. We investigate significance of low frequency-based acoustic features, in particular, Generalized Morse Wavelet (GMW)-based features for ADD. To that effect, we estimated Continuous Wavelet Transform (CWT), having GMW as mother wavelet function [4]. We used Morse wavelet to form a scalogram before testing them on the CNN classifier. Motivated by encouraging results by Mel spectrogram features, we explored other lower frequency-based features, in particular, GMW-based features for ADD task. Relative significance of lower frequency region than full band (w.r.t. Shannon sampling paradigm) is also illustrated in this study.

Both basic types of data fusion strategies namely, feature-level fusion and score-level fusion were examined in order to investigate possible *complementary* information captured by individual systems of both the fusions. After release of the Fake or Real (FoR) dataset, much progress has been made on machine learning-based approaches, and also a few deep learning-based approaches have been explored. In [5], the authors explore machine learning-based approach, such as Long Short-Term Memory (LSTM), and VGG-16 for ADD task on FoR dataset. In [6], the authors proposed the use of baseline Mel Frequency Cepstral Coefficients (MFCC), with a variety of models, such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost), for ADD task. However, the feature-based approach still remains unexplored. In this study, authors propose GMW-based features fused with Mel spectrogram for ADD task. In part, this overcomes the practical problem of choosing the appropriate mother wavelet function because generalized Morse wavelet function is a super-family of different analytic wavelets [7]. Also unlike classic analytical wavelet (which is approximately analytic and thus resulting in spectral leakage at lower frequency), such as Morlet, Morse wavelet shows exact systematic behaviour (i.e., no spectral leakage).

The Morse wavelet-based scalogram and Mel-spectrogram both are known to have a good resolution in the lower frequency regions. However, the scalogram as shown in Fig. 1 (a) has much higher frequency resolution for frequencies below 0.1 kHz (i.e., 100 Hz) and it can be observed from the scalogram that there is hardly any significant energy in that region. So the benefit of using Morse scalogram because of its excellent frequency resolution property in that region is not well utilized for ADD task. However, for the Mel spectrogram, even though the frequency bins are spaced in a Mel scale, the Mel spectrogram is able to emphasize on frequency regions beyond 100 Hz as well. In Fig. 1 (b), we can note that the harmonic structure for spectrogram is less clear than Mel spectrogram and Morse wavelet spectrograms. Similar observations can be made in Fig. 1 (c) (Panel I), has

abundantly blurred pitch source harmonics (i.e., kF_0, $k \in Z$) structure, as compared to Fig. 1 (c) (Panel II), which has sharp harmonic structure, resulting into clear classification while using Mel spectrogram for ADD task. According to authors best knowledge, this is the first study that employes and explores wavelet based features for ADD task. The remaining paper is organized as follows: Section II provides the details about lower frequency-based features. Section III presents information about experimental setup used for this study. Section IV presents experimental results and the related discussion. Section V provides summary, conclusion and the future potential research directions for this study.

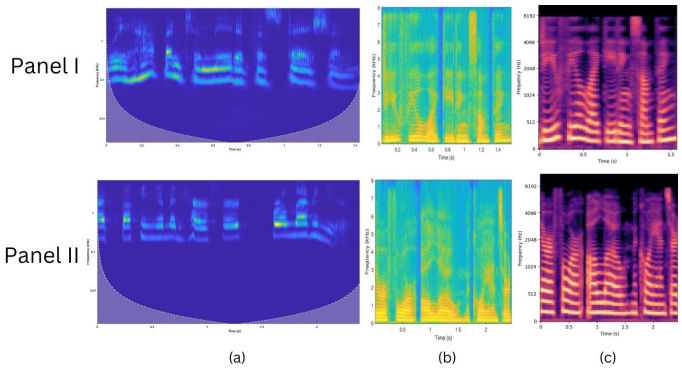

Fig. 1: Panel I (fake signal), and Panel II (real signal): (a) Morse wavelet spectrogram, (b) Spectrogram, and (c) Mel Spectrogram, respectively.

II. Lower Frequency-Based Approach For Feature Extraction

A. Mel Spectrogram

Due to mathematical structure of Mel scale, Mel spectrogram of a speech is known to emphasize lower frequency region more as compared to higher frequency region [8]. The spectrogram of an speech signal ($m(t_0)$) can be defined as magnitude square of inner product. In particular, i.e.,

$$|< \mathrm{m}(t_0), g_{u,\zeta}(t_0) >|^2 = \left| \int_{-\infty}^{\infty} m(t_0) g(t_0 - u) e^{-j\zeta t_0} dt_0 \right|^2, \tag{1}$$

where $g_{u,\zeta}(t_0)$ is called time-frequency atoms. We extracted Mel spectrograms by 30 *ms* analysis window length, 20 *ms* window overlap, 32 number of sub-band filters, and taking normalized window. For such scalogram-based features, we were motivated to use Morse wavelet features following its success to various audio-based tasks [9].

B. Continuous Wavelet Transform (CWT)

After the success of MFCC and Mel spectrogram features [5], [6], it is important to investigate the significance of low frequency regions for the ADD. The Mel frequency spectrogram and MFCC features follow the Mel scale, which follows the perceptual scale of hearing by humans. The Mel scale given by :

$$Mel\_Scale = 1127 \ln \left(1 + \frac{f}{700}\right), \tag{2}$$

where f is in Hz and hence, has higher frequency resolution in the lower frequency regions, in comparison high frequency regions. To further investigate the significance of the low frequency regions for the ADD, for this study, we propose the use of CWT-based time-frequency representation, popular as the *scalogram*. CWT-based representations are well known for *Multi-Resolution Analysis (MRA)*. In order to achieve this, we used a CWT-based method that finally produced high frequency resolution in low frequency limits. For a signal $m(t)$, CWT is given by :

$$\begin{aligned} W_m(a,b) &=< m(t_0), \psi_{a_0,b_0}(t_0) >, \\ &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} m(t_0) \psi^* \left(\frac{t_0 - b_0}{a_0} \right) dt_0, \end{aligned} \tag{3}$$

where $< \cdot, \cdot >$ is inner product operation, which is used to calculate wavelet coefficients, and * represents complex conjugate. In Eq. (3), parameter a_0 is known as the scaleparameter, whereas parameter b_0 is known as translation parameter. A scalogram is a graphic representation of energy of CWT coefficients (i.e., $| < m(t_0), \psi_{a_0,b_0}(t_0) > |^2$). In this study, we computed CWT coefficients by taking into account the Morse wavelet. The word wavelet means a short duration wave. Alternatively, not every short wave is called a wavelet. A wavelet $\psi(t_0) \in L^2(R)$ (i.e., Hilbert space for signal with finite energy and Lebesgne integrable functions) weekly satisfies the condition, represented by :

$$C_\psi = \int_0^{\infty} \frac{|\Psi(\omega_0)|^2}{\omega_0} d\omega_0 < \infty, \tag{4}$$

In Eq. 4, $\Psi(\omega_0)$ represents Fourier transform of mother wavelet function, $\psi(t_0)$. This proves $\Psi(\omega_0)$ drops its value to zero as fast as $\omega_0 \to 0$. Furthermore, to validate Eq. (4), we are required to impose $\Psi(0) = 0$, which is equivalent to :

$$\Psi(\omega_0)|_{\omega_0=0} \equiv \int_{-\infty}^{\infty} \psi(t_0) dt_0 = 0. \tag{5}$$

Further, for task of energy normalization, we consider wavelets satisfying $\|\psi(t_0)\| = 1$, i.e., $\int_{-\infty}^{\infty} |\psi(t_0)|^2 dt_0 = 1$.

1) Proposed GMW-based Features: Systematic wavelets exist in several types, such as complex Shannon, log-normal, Gaussian derivative, Cauchy, and Morlet wavelets, and are mostly noted in the frequency-domain [10]–[13]. GMW represent the family of systematic wavelets, which addresses the problem of choosing a specific type of analytic wavelet [7], [14], [15]. GMWs represent a set of eigenvectors w.r.t. to a time-frequency localization operator, whose details can be studied from [16]. From the set of eigenvectors, the largest eigenvector is considered, which makes the 0^{th} GMW. The largest eigenvector is known to enable *optimal energy concentration* properties [16], [17]. To that effect, we represent the 0^{th} GMW as the Morse wavelet. It is an analytic wavelet and is expressed mathematically as [7], [14], [17]:

$$\Psi_{\beta,\gamma}(\omega_0) = \int_{-\infty}^{+\infty} \psi_{\beta,\gamma}(t_0) e^{-j\omega_0 t_0} dt_0, \tag{6}$$

where the parameters β and γ let degree of freedom additionally and make GMWs form a *family* of analytic wavelets [15]. This degree of freedom is known to establish the wide range of behaviour of the GMWs depending on the parameters β and γ. To that effect, Eq. (6) can be said to be a general approach to construct an *exactly* analytic wavelet [15]. Furthermore, Eq. (6) ensures that the wavelet satisfies the admissibility condition, which require the wavelet to have finite energy and zero-mean, i.e.,

$$\int_{-\infty}^{+\infty} |\psi_{\beta,\gamma}(t_0)|^2 dt_0 = 1. \tag{7}$$

For analysis purposes, the dimensionless derivatives of the Morse wavelet are considered [15]. In particular, its 2^{nd} order derivative is denoted by $P^2_{\beta,\gamma}$, and is called the *wavelet duration* or *inverse bandwidth*. The significance of the 2^{nd}-order derivative is discussed in detail in the Appendix. The quantity $P^2_{\beta,\gamma}/2\pi$ sets the number of oscillations in the wavelet, where $P^2_{\beta,\gamma}$ is given as [15]:

$$P^2_{\beta,\gamma} \equiv -\frac{\omega^2_{0\beta,\gamma}\Psi''_{\beta,\gamma}(\omega_{0\beta,\gamma})}{\Psi_{\beta,\gamma}(\omega_{0\beta,\gamma})} = \beta\gamma. \tag{8}$$

Furthermore, when $P^2_{\beta,\gamma}$ is fixed, and γ is varied, the wavelet behaviour corresponds to different families of wavelets. Hence, the parameter γ is also termed as the *family*. Further, the parameter γ is also known to be related to area of Heisenberg's box in the time-frequency plane [7]. In particular, $\gamma = 3$ transforms to the *'Airy family'*, where the Morse wavelet behaves similarly to the conventional Morlet wavelet. Interestingly, the conventional Morlet wavelet has significance in the human perception process for both hearing and vision [18]. Human hearing mechanism is historically known to be basis for the developing various state-of-the-art features in speech-related pattern recognition problems. In particular, the well known MFCC features are derived by Mel frequency wrapping, which is truely motivated by the Weber's law of human perception, also developed using the perception mechanism of human hearing [19]. Likewise, the Mel scale is also known as the perceptual scale of hearing. This motivated us to explore the perceptual scales and hence, the low frequency regions for the ADD. In this paper, we consider the Morse wavelet with $\gamma = 3$, instead of the conventional Morlet wavelet, because the coventional Morlet wavelet can be observed to defer from the analyticity under special cases, however, the Morlet-like wavelet obtained by taking $\gamma = 3$ in Morse wavelet, shows *strictly* analytic behaviour, as shown in the Fig. 2. In particular, the Morlet wavelet, as observed in Fig. 2 (b), shows a *spectral leakage* in the negative frequencies as shown by its Wigner Ville distribution. On the contrary, as observed in Fig. 2 (d), upon analysis of the Morse wavelet at $\gamma = 3$, it can be observed that there is no spectral leakage in the negative frequency regions of the GMW. This indicates that strict analytic properties are exhibited by the Morse wavelet, which is a characteristic of wavelets that satisfy the Cauchy-Riemann equations of differentiablity. This means that the wavelet function $\Psi(\omega)$ equals zero for $\omega < 0$. The Morlet wavelet, on the other hand, does not exhibit strict analytic properties, as its wavelet function $\Psi(\omega)$ has non-zero values for $\omega < 0$. This makes the Morse wavelet a preferred choice in applications, where the presence of negative frequency components can lead to inaccuracies in the analysis, and in particular, the ADD task discussed in this paper. We believe that this remarkable exact analytic properties of morse wavelets (i.e., $\psi(\omega) = 0$ for $\omega < 0$) may help to efficiently capture class-specific attributes (e.g. real *vs.* deepfake) of given signal that are present neat lower frequency regions, whereas for approximately analytic wavelets (such as Morlets), the energy in their lower frequency region may experience blur primarily due to spectral leakage of energy in the negative frequencies [19]. Fig. 3 represents the flowchart of process carried out while classification of fake and real audios.

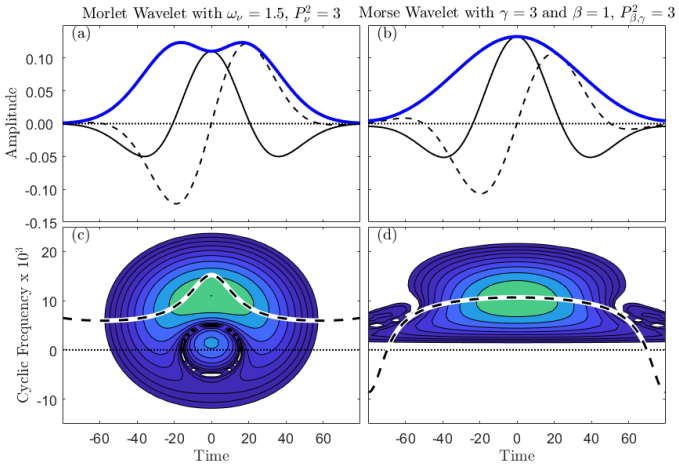

Fig. 2: Analysis of spectral leakage (a) Morlet wavelet, *vs.* (b) Morse wavelet, and (c), (d) their respective Wigner-Ville distributions. After [15], [19].

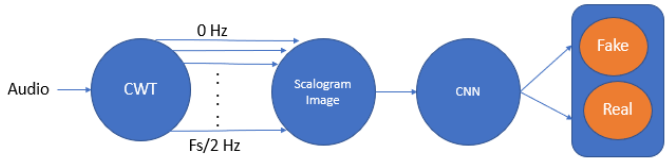

Fig. 3: Functional block diagram of proposed ADD system.

III. Experimental Setup

A. Dataset Used

1) Fake or Real (FoR) Dataset: It is one of the most well known, statistically meaningful, and freely available dataset for ADD tasks[20]. It contains around 87,000 fake utterances as well as 1,11,000 real utterances, which were generated by more than 1200 speakers. Further, FoR has more number of training files as compared to ASVSpoof 2019 dataset and hence, FoR is used in this paper. The large amount of data makes the model train properly, thereby leading to proper evaluation of model. The main reason to use the FoR dataset instead of others is that it includes speech produced by state-of-the-art voice synthesis algorithms. A high number of files in the dataset can also

classify points into different classes without overfitting. For dataset have been released in 4 versions. FoR-original is the version, which has original collected utterances. FoR-norm is the version of dataset that has samples which were collected after performing certain set of prepossessing on the original audio. FoR-2second contains the samples truncated to 2 second and at last for-rerecorded consist of all utterance which were recorded again from the original audio with gadgets like speaker and microphone. Released in the early half of 2021, it makes it one of the most recently released datasets. The authors in [21], used ASVspoof dataset which consists of total 1,25,000 samples, compared to which FoR consist of 1,95,541 utterances of total. We firmly believe that, more the number of data samples used for training dataset, more accurately the model gets trained. As a result to which authors chooses FoR dataset over ASVspoof dataset. Because the *FoR-norm* version of the FoR dataset is normalized and preprocessed, the authors have been using it for all experiments in this manuscript's study. This is because we feel it to be the best option for more accurate and thorough model training. The gender ratio in the FoR-norm version is clearly more balanced than in the FoR-original dataset, which is another incentive to utilize it. It has been divided in training (77.73%), testing (6.68%), and validation (15.58%) sets, to make up the dataset. The sampling rate of speech data is 16 kHz.

B. Pattern Classifier Used

The CNN classifier used in this work consists of a Fully-Connected (FC) layer after two consecutive 2D-convolutional layers. The CNN is provided with an input feature dimension of 512 × 512. Using a 7 x 7 kernel, the first Convolutional Neural Layer (CNL) contains 3 input channels ($I_{channels}$) and 8 output channels ($O_{channels}$). The second CNL has a 3 × 3 kernel, 8 $I_{channels}$, and 16 $O_{channels}$. The last CNL has a 3 x 3 kernel, 16 $I_{channels}$, and 32 $O_{channels}$. Every CNL has padding and stride values of 1. A max-pool layer with a 3 × 3 kernel and a stride of 3 is used to reduce dimensions. The activation function is the Rectified Linear Unit (ReLU). The choice for optimizer was Adam. The Binary cross-entropy Loss Function (BceLF) is chosen as the loss metric. This architecture and configuration are designed for tasks involving input images of size 512 × 512, demonstrating a multi-layered approach for hierarchical feature extraction and classification. Learning rate was defined as 0.001.

C. Performance Metrics

1) Precision: It is implemented to measure how much accurately the model works on positive samples, which also predicts how many positive samples are correctly classified.

2) Recall: Alternet to Precision recall works somewhat similar to precision. The main difference is, Recall measures the percentage of actual positive cases that were accurately anticipated to be positive.

3) F1-Score: It balances Precision and Recall scores. It can simply be defined as harmonic mean of precision and recall.

IV. Experimental Results

This Section examines and proposes ML / DL-based experiments performed for measuring the accurate working of proposed feature vector in combination with pattern classifier. Table I presents the comparison of results w.r.t. several features used on CNN classifier. The Mel spectrogram, which capture information about how frequency region gave us an significant result of 95.07 % accuracy, which is 17.24 % more than spectrogram. This motivated us to explore other features having good resolution in lower frequency region. The proposed GMW features gave 79.56 % accuracy, which is 1.73 % higher than spectrogram. Since recall is higher in Morse wavelet and spectrogram, it indicates that false negative samples are much less in comparison with true negative. Since Mel spectrogram gave high precision, the number of false positive samples are less compared to the Morse wavelet and spectrogram. Low precision in Morse wavelet and spectrogram indicates that these two features are not able to predict real speech properly, whereas Mel spectrogram is not able to predict fake speech precisely as compared with real speech.

TABLE I: Results *w.r.t.* Different Explored Features and CNN Classifier

Features	Accuracy (in %)	Precision	Recall	F1-Score
Spectrogram	77.83	68.85	99.77	76.94
Morse wavelet	**79.56**	74.52	88.38	79.46
Mel Spectrogram	**95.07**	97.52	92.27	95.06

A. Results using Feature-Level Fusion

We first extracted two feature sets and fused them by concatenat like a stack, which resulted in a new feature vector of dimension 1024 x 512 pixels per frame. Table II represents results *w.r.t.* feature-level fusion. By performing feature-level fusion of Mel Spectrogram and Morse wavelet, we achieved an accuracy of 97.00 %, which is 1.93 % higher than Mel spectrogram alone. By performing feature-level fusion, we are also able to get Equal Error Rate (EER) as low as 2.99 %. On the other hand, the fusion of Mel spectrogram and Spectrogram resulted in an accuracy of 90.22 %.

TABLE II: Feature-Level Fusion on CNN Classifier

Feature 1	Feature 2	Accuracy (in %)	EER
Morse wavelet	Spectrogram	79.08	20.91
Mel Spectrogram	Spectrogram	90.22	9.75
Morse wavelet	Mel Spectrogram	**97.00**	**2.99**

B. Results using Score-Level Fusion

After increase in ≈ 3 % accuracy by performing feature-level fusion, the authors explored the potential of score-level fusion for two features. In this type of fusion, we fused the scores gained by each model while testing, which were taken in ratio of α and $(1-\alpha)$, where $\alpha \in [0,1]$, i.e., $\alpha S_{Morse} + (1+\alpha)S_{spectrogram}$, where S_{Morse}, and $S_{spectrogram}$ are scores for wavelet, and spectrogram, repressively. Fig. 4 shows that Mel Spectrogram and spectrogram when fused at the score-level gives a bit less accuracy than the accuracy gained by

fusing Morse wavelet and spectrogram. However, score-fusion of Morse wavelet and spectrogram keeps on performing poorly and thus, this fusion not pursed further for ADD. This findings is in agreement with other studies in pattern recognition literature, where various data fusion strategies, such as, signal-level, feature-level, classifier-level, and score level may give different inferences w.r.t.complementary information captured by various individual pattern recognition systems. In particular, in our study, we found that feature-level fusion captures better complementary information then score-level fusion.

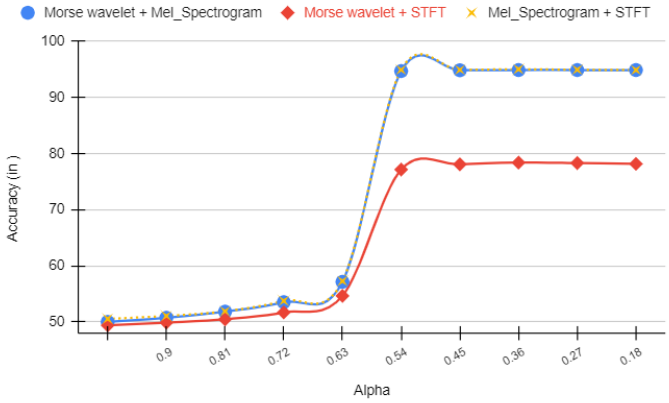

Fig. 4: Score-Level Fusion of Features.

C. Discussion

Mel spectrogram as well as GMW features emphasize lower frequency region, however, both are not giving the same results, which is due to the difference in relative joint time-frequency resolution captured by spectrogram *vs.* CWT as per Heisenberg's uncertainty principle in signal processing framework [22]. The amount of frequency resolution of CWT is comparatively very high than Mel spectrogram. In order to analyze their relative feature discrimination power, we perform feature-level fusion of both the features [23]. Even though Morse wavelet is not performing good, giving us an accuracy of 79.56 %, when fused with the Mel spectrogram, which is initially giving accuracy of 95.07 %, the accuracy jump to 97 %, indicating that the Mel spectrogram is also able to capture characteristics of a few higher frequency regions, which on the other hand, Morse wavelet is unable to capture alone. In order to detect deepfake attack, we need frequency region from all possible sub-bands, i.e., from 0 to $Fs/2$, however, the emphasis should be laid on the lower frequency region as captured by the Mel spectrogram. On the other hand, spectrogram (which employ Windowed Fourier Transform (WFT) or Short-Time Fourier reasnform (STFT)) has *constant* resolution in each frequency band (in the entire time-frequency panel), i.e., it captures information from all frequency range. Spectrogram not emphasizing to lower frequency region particularly, resulting in low accuracy performance than other lower frequency resolution-based features. Fusion of other features with spectrogram also leads to lower accuracy than individual lower frequency resolution-based feature sets.

V. Summary and Conclusions

In this study, we employed features based on low frequency regions, namely, GMW and Mel spectrogram for ADD task. In particular, we employed Morse wavelet and Mel spectrogram with CNN as pattern classifier, which gave relatively better results. Further, feature-level fusion, we are able to obtain even more higher accuracy than the individual features. However, the proposed approach requires high time and computational storage. Furthermore, many other wavelet-based approaches based on lower frequency resolution can be explored for ADD task. Future plans includes experiments to explore more low frequency resolution-based features, as well as other pattern classifiers for ADD task. Historically, the discrete time implementation of CWT involves discrete wavelet bases $\psi_{j,k}(t_0) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t_0 - 2^j \cdot k}{2^j}\right)$ that are related to multi-resolution analysis (MRA) and its implementation and its implementation via Quadrature Mirror Filterbank (QMF) involving cascade of lowpass and highpass filters relating several properties of wavelets, such as vanishing moments, smoothness, regularity, and support size [24]. Developing this theoretical as well as practical viewpoint for GMW remains an open research problem. Our findings indicate that lower frequency content in deepfake is discriminating in nature for ADD, which may be attributed to Instantaneous Frequencies (IFs) in lower frequency region. Thus, exploring Morse Wavelets for IF estimation in the classic Delprat's IF estimation using ridge analysis remains another challenging open research problem.

Appendix

The k^{th} order moment of the demodulated morse wavelet by the peak frequency ω_ψ can be written as [19] :

$$m_{k;\psi} = \int_{-\infty}^{+\infty} t^k \psi(t) e^{-i\omega_\psi t} dt. \tag{9}$$

As per the relationship between moments in time-domain and derivatives in frequency-domain,

$$\frac{m_{k;\psi}}{m_{0;\psi}} = i^k \frac{\psi^{(k)}(\omega_\psi)}{\tilde{\Psi}(\omega_\psi)}, \tag{10}$$

where $\tilde{\Psi}(\omega_\psi)$ is the wavelet's *dimensionless derivatives* at $\omega_0 = \omega_\psi$, and defined as [15], [17]:

$$\tilde{\Psi}_k(\omega_0) \equiv \omega_0^k \frac{\Psi^{(k)}(\omega_0)}{\Psi(\omega_0)}, \tag{11}$$

where k as superscript denotes the k^{th} order derivative, and $\tilde{\Psi}(\omega_\psi)$ is nothing but $\tilde{\Psi}_k(\omega_0)$ evaluated at $\omega_0 = \omega_\psi$. Therefore,

$$\frac{m_{k;\psi}}{m_{0;\psi}} = i^k \frac{\psi^{(k)}(\omega_\psi)}{\tilde{\Psi}(\omega_\psi)} = i^k \frac{\tilde{\Psi} k(\omega_\psi)}{\omega_{0\,\psi}^k}. \tag{12}$$

Since the wavelet satisfies the weak admissibility condition, its mean is zero ,i.e., $m_{1;\psi} = 0$. Therefore, the higher-order demodulate moments can be normalized by the 2^{nd} moment

and hence, the dimensionless measure using the 2^{nd} moment is defined as

$$P_{\beta,\gamma} = \omega_\psi \sqrt{\frac{m_{2;\psi}}{m_{0;\psi}}} = \sqrt{|\tilde{\Psi}_2(\omega_\psi)|}, \tag{13}$$

where $\tilde{\Psi}_2(\omega_\psi) = -\beta\gamma$ (as represented in [15]).

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Priyanka Gupta, LNMIIT, Jaipur for her help during this research work. Further, they acknowledge partial support from Ministry of Electronics and Information Technology (MeitY), Grant ID: 11(1)2022-HCC.(TDIL).

REFERENCES

[1] S. Maddocks, "'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes," *Porn Studies*, vol. 7, no. 4, pp. 415–423, 2020.

[2] A. R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, "A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions," *Engineering Applications of Artificial Intelligence*, vol. 106, pp. 104–456, 2021.

[3] A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, "Privacy of web browsers: A challenge in digital forensics," in *Genetic and Evolutionary Computing: Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computing, October 21-23, 2021, Jilin, China 14*, Springer, 2022, pp. 493–504.

[4] P. Gupta, S. Gupta, and H. Patil, "Voice liveness detection using bump wavelet with cnn," in *9th International Conference on Pattern Recognition and Machine Intelligence*, 2021.

[5] 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.

[6] 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*, pp. 1–12, 2021.

[7] J. M. Lilly and S. C. Olhede, "Generalized morse wavelets as a superfamily of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6036–6041, 2012.

[8] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, Vol 3, 2022.

[9] E. A. Martinez-Ríos, R. Bustamante-Bello, S. Navarro-Tuch, and H. Perez-Meana, "Applications of the generalized morse wavelets: A review," *IEEE Access*, Vol 3, 2022.

[10] M. Holschneider, "Wavelets : An Analysis Tool," *Oxford Science Publications*, 1995.

[11] S. G. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 2^{nd} Edition, 1999.

[12] H. Knutsson, C.-F. Westin, and G. Granlund, "Local multiscale frequency and bandwidth estimation," in *Proceedings of* 1^{st} *International Conference on Image Processing (ICIP)*, vol. 1, 13-16 Nov. 1994, pp. 36–40.

[13] X. Zhu and J. Kim, "Application of analytic wavelet transform to analysis of highly impulsive noises," *Journal of Sound and Vibration*, vol. 294, no. 4-5, pp. 841–855, 2006.

[14] S. C. Olhede and A. T. Walden, "Generalized Morse Wavelets," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2661–2670, 2002.

[15] Lilly, Jonathan M. and Olhede, Sofia C., "Higher-order properties of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 146–160, 2008.

[16] I. Daubechies and T. Paul, "Time-frequency localization operators-a geometric phase space approach: II. The use of dilations," *Inverse Problems*, vol. 4, no. 3, p. 661, 1988.

[17] J. M. Lilly and S. C. Olhede, "On the analytic wavelet transform," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 4135–4156, 2010.

[18] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[19] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *Computer Speech & Language*, vol. 84, p. 101 571, 2024.

[20] 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.

[21] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," *arXiv preprint arXiv:2306.01428*, 2023.

[22] P. Busch, T. Heinonen, and P. Lahti, "Heisenberg's uncertainty principle," *Physics Reports*, vol. 452, no. 6, pp. 155–176, 2007.

[23] R. A. Rasool, "Feature-level vs. score-level fusion in the human identification system," *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–10, 2021.

[24] S. G. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$," *Transactions of the American Mathematical Society*, vol. 315, no. 1, pp. 69–87, 1989.