A Frequency Bin Analysis of Distinctive Ranges Between Human and Deepfake Generated Voices

Anonymous Double-Blind IJCNN 2024 Submission

***Abstract*—Deepfake technology has advanced rapidly in recent years. The widespread availability of deepfake audio technology has raised concerns about its potential misuse for malicious purposes, and a need for more robust countermeasure systems is becoming ever more important. Here we analyse the differences between human and deepfake audio and introduce a novel audio pre-processing approach. Our analysis aims to show the specific locations in the frequency spectrum where these artefacts and distinctions between human and deepfake audio can be found. Our approach emphasises specific frequency ranges that we show are transferable across synthetic speech datasets. In doing so, we explore the use of a bespoke filter bank derived from our analysis of the WaveFake dataset to exploit commonalities across algorithms. Our filter bank was constructed based on a frequency bin analysis of the WaveFake dataset, we apply this filter bank to adjust gain/attenuation to improve the effective signal-to-noise ratio, doing so we reduce the similarities while accentuating differences. We then take a baseline performing model and experiment with improving the performance using these frequency ranges to show where these artefacts lie and if this knowledge is transferable across mel-spectrum algorithms. We show that there exist exploitable commonalities between deepfake voice generation methods that generate audio in the mel-spectrum and that artefacts are left behind in similar frequency regions. Our approach is evaluated on the ASVSpoof 2019 Logical Access dataset of which the test set contains unseen generative methods to test the efficacy of our filter bank approach and transferability. Our experiments show that there is enhanced classification performance to be gained from utilizing these transferable frequency bands where there are more artefacts and distinctions. Our highest-performing model provided a 14.75% improvement in Equal Error Rate against our baseline model.**

***Index Terms*—Automatic Speaker Verification (ASV) anti-spoofing, Deepfake, synthetic speech detection, SEResNet, filter bank**

I. Introduction

Deepfake technology, which involves manipulating images, audio, video or text content to create convincing fake representations [1], has advanced rapidly in recent years. The widespread availability of deepfake technology has raised concerns about its potential misuse for malicious purposes, such as impersonating individuals in voice-based phishing scams, also known as "vishing". The case of a UK energy CEO in 2019 being defrauded out of 243,000 dollars through a deepfake phone call [2] highlights the urgent need for effective automatic speaker verification (ASV) countermeasures to combat deepfake voice misuse. Within the past few years, the ability of a human to be able to reliably tell the difference between genuine and deepfake voice has come into question, with models such as the VITS Text-To-Speech (TTS) and voice conversion (VC) models producing subjective naturalness mean opinion scores equal to that of genuine human voice [3].

However, there are still telltale artefacts left behind by the generation or conversion algorithm. Such artefacts can be very difficult for humans to notice due to how humans perceive the volume of a given frequency. The responsiveness of the human auditory system is known to follow the Weber-Fechner law which states that the intensity of a sensation is proportional to the logarithm of the intensity of the stimulus [4]. This means humans can much more easily detect differences at lower frequencies than at higher frequencies. This poses a problem if the telltale signs of deepfake audio are at higher frequencies.

Much research has been conducted predominantly concentrating on conventional audio features that focus on the lower frequencies of spoken audio, usually by using a logarithmic or mel-scaled y-axis. Examples include log-mel spectrograms for dual audio spoof and video deepfake detection [5], prosody and speaker embeddings produced from Mel Frequency Cepstral Coefficients (MFCC) [6] as well as the vast usage of the Constant Q Cepstral Coefficient (CQCC) and Constant Q Transform (CQT) [7]. However, such scaling of the y-axis of the spectrogram diminishes the information in the higher frequencies as a trade-off to enhance the fidelity of the lower frequencies. Others have explored the use of features specifically accentuating high-frequency discriminative patterns and audio artefacts [8] [9], with Inverse Mel Frequency Cepstral Coefficients (IMFCC) showing promising performance for unseen generative algorithms [10] as it accentuates the higher frequencies instead of the lower ones. Furthermore, the authors of the WaveFake dataset have remarked upon substantial distinctions in the higher frequencies between authentic human audio and synthetically generated audio, particularly the generative outputs of the MelGAN and WaveGlow vocoders [11]. This can be explained by the fact that many neural vocoders synthesise a waveform from the mel-spectrum [12], by producing a waveform from the mel-spectrum more fidelity is given to the lower frequencies but as a trade-off less fidelity is given to the higher ones. Nevertheless, there exists a research gap in identifying the specific locations where these artefacts and distinctions between human and deepfake audio can be found.

In this work, we explore the differences between human and deepfake audio through an analysis of the WaveFake

dataset [11]. This analysis aims to show which frequency bands contain the greatest amount of difference between human and deepfake audio which is crucial for the effective detection of deepfake speech. We employ the application of a custom-designed filter bank, derived from the frequency band-wise analysis of the WaveFake dataset. Our exploration involves an examination of frequency bands within the WaveFake dataset, revealing pronounced differences in the higher-frequency regions compared to the lower-frequency regions. To provide a validation of these findings, we conduct a dip-sample confirmation by randomly selecting and comparing spectrograms of deepfake voice with those of authentic human voice. Additionally, we employ a baseline performance model to investigate the extent to which amplifying frequency bands generalize from one dataset to another, aiming to enhance overall performance. Ultimately, this shows that there exist commonalities between mel-spectrum generative methods, in that most of the differences between human and deepfake audio exist within similar frequency ranges.

The contributions of this paper include: 1) Analysis of Frequency Bands Across Generative Vocoders: Our analysis uncovers shared patterns in the synthesis of deepfake audio across various generative vocoders. Through our analysis of the WaveFake dataset, derived from LJSpeech and JSUT (Japanese speech corpus of Saruwatari-lab., University of Tokyo), we identify common frequency regions amongst different deepfake generative vocoders that are more different (measured in Root Mean Square Error (RmSE)) when compared with human audio. This finding suggests that these shared frequency regions can be harnessed to improve the classification results of deepfake audio across a spectrum of generative techniques and datasets. We evaluate the resilience and generalization of our approach by testing it on the ASVspoof 2019 LA dataset. The test set of this dataset comprises unseen generative methods, allowing us to assess the model's performance in detecting synthetic audio content generated by unseen mel-spectrum generation methods. Our goal is to extend the robustness of audio classification models and contribute to the advancement of effective strategies for identifying synthetic audio across a spectrum of generative methods. 2). A demonstration of enhanced classification performance when utilised as an anti-spoofing countermeasure: In this study, we introduce a novel pre-processing approach that exploits these frequency regions that are shown to be common across mel-spectrum generative vocoders. By leveraging the shared characteristics in WaveFake, we achieve a notable improvement in Equal Error Rate (EER) performance from 10.334 to 8.810. This enhancement represents a 14.75% improvement in EER against our baseline model and a trend of improvement by applying this novel pre-processing step alone.

II. Methodology and Analysis

In this section, we explore the methodology and analysis of our investigation into the distinctions between deepfake and human audio. Initially, we hypothesize that these differences are not uniformly distributed across frequencies, challenging the null hypothesis that posits an even spread. Specifically, we propose that higher frequencies, potentially imperceptible to the human ear, harbour more pronounced disparities. Our ensuing analysis substantiates this hypothesis, revealing a concentration of differences in the higher frequency ranges.

A. Band-Wise Analysis

For our analysis, we used the WaveFake dataset, which consists of audio generated using 7 different deep-neural generative vocoders sampled at 22.1kHz. The dataset contains English and Japanese from the LJSpeech and JSUT datasets respectively using the following vocoders: HifiGan, Fullband MelGan, MelGan, MelGan Large, Multiband MelGan, Parallel WaveGan and WaveGlow. We used the English LJSpeech portion for our analysis [13]. the authors produced the WaveFake dataset by running the LJSpeech authentic speech dataset through the 7 vocoders to produce a deepfake version of the input with the telltale signs of the vocoder that was used to generate it. The purpose is that this can be used for investigating the specific signs that each vocoder leaves behind through the generative process. Most importantly, this dataset controls for the speaker as the same speaker is used throughout and allows for comparison with the original LJSpeech audio. Controlling for the speaker therefore allows for vocoder-specific patterns to be observed, any differences are a result of the vocoder as the speaker is a controlled factor, with the vocoder being a dependent variable. For the audio of each generative vocoder in the Wavefake dataset, we generated Short-time Fourier Transform (STFT) spectrograms with an ordinal linearly spaced y-axis. From this, we split the spectrogram into 40 frequency band sections. Each of these frequency bands from each piece of audio was then compared with the original genuine human audio from the LJSpeech dataset. The frequency band sections were compared element-wise using the Root Mean Squared Error (RMSE) of the pixel values between the spectrogram images. The full results can be seen in Figure. 1 which shows the RMSE between human and deepfake for each algorithm at each frequency bin. To analyze the WaveFake audio, we employed a filter bank, a signal processing technique used to decompose a signal into its frequency components. The filter bank consists of a series of bandpass filters, each designed to isolate a specific frequency range from the input audio signal.

The results of the analysis in Figure. 1 show multiple patterns that remain true for all generative vocoders. There exists a distinct lack of difference between the lower frequencies (0 - 3kHz), predominantly at around 2kHz. This shows that neural vocoders generating audio in the mel-spectrum are very effective at recreating the lower frequencies. However, as the frequency increases beyond around 3kHz the amount of difference between human and deepfake increases, this shows that there is more discriminatory information in the higher frequencies.

Additionally, we analyzed the audio from various vocoders in the WaveFake dataset. In Figure. 2 we can see in the human spectrogram (left) that the fundamental frequency (F0) and

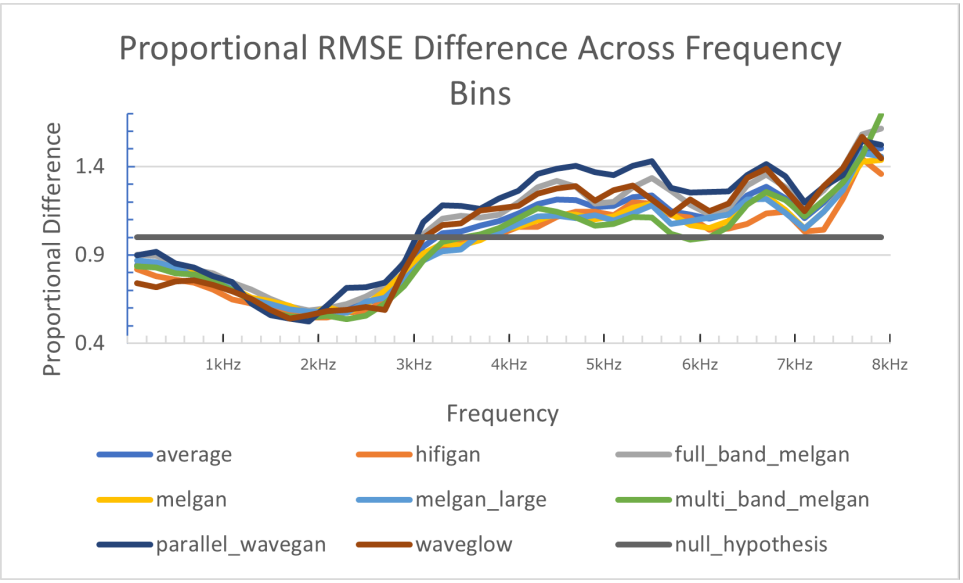

Fig. 1. Comparison of Algorithmic Differences Across 40 Frequency Bins. 200Hz intervals, 0 - 8000Hz. Bin 1 = 0-200Hz, Bin 2 = 201-400Hz, etc.

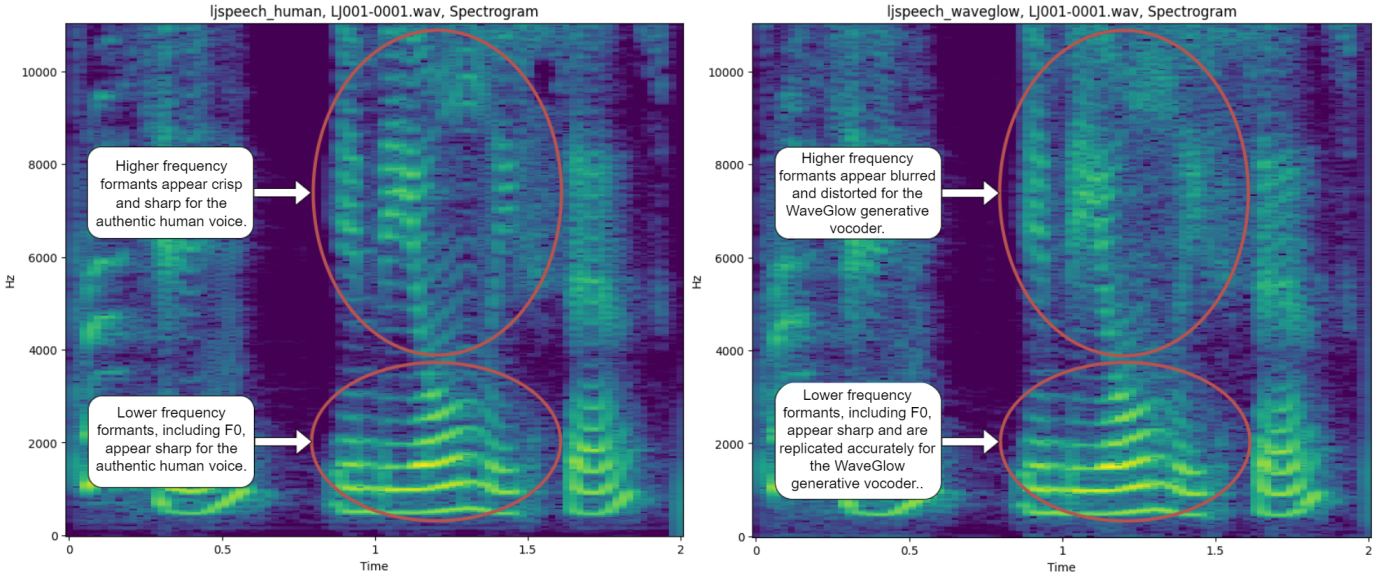

Fig. 2. Spectrogram comparison between Human (left) and WaveGlow (right).

successive formants are crisp and represented accurately as expected. However, the spectrogram on the right (WaveGlow) in Figure. 2 shows that while the F0 and low-frequency formants such as F1 and F2 appear to be recreated accurately, subsequent formants in the higher frequencies are not replicated properly and appear blurred and distorted. This observation is in line with what we saw with our frequency band analysis in Figure. 1. There appear to be fewer differences between real and fake in the lower frequencies as the audio is replicated more accurately, whereas there are more differences in the higher frequencies. This pattern can be observed across all of the vocoders in the WaveFake dataset and be explained by the fact that all of the vocoders synthesise a waveform from the mel-spectrum. Finally, as previously mentioned these high-frequency distortions may be difficult for humans to perceive due to the human perception of stimulus scaling logarithmically.

B. Filter Bank

Using the per-band analysis results, we applied specific adjustments to the target audio on a per-frequency band basis. This involved the amplification or attenuation of the audio tailored to the frequency bands identified in the analysis shown in Figure. 1. By amplifying and attenuating bands, we aim to amplify differences between human and deepfake voices, boosting discriminative information while selectively reducing

specific frequency ranges where the audio is similar. This novel technique aims to enhance the signal-to-noise ratio of the audio signal by attenuating frequency bands that are more similar while amplifying bands that are more different. As the WaveFake dataset is sampled at 22100Hz and the ASVSpoof 2019 LA dataset is sampled at 16000Hz, we re-sampled the WaveFake audio using the Python library Librosa so that they were at the same sampling rate. The audio was re-sampled and not down-sampled to preserve the information within the audio and prevent aliasing. For our filter bank implementations, we utilized a Butterworth bandpass filter to partition the audio signal into distinct frequency bands expressed in Equation. 1. The Butterworth filter was chosen for its desirable characteristics, including a smooth frequency response. The design parameters of the Butterworth filter, such as the lower and higher cutoff frequencies, were determined based on the specific frequency bands defined for our application, in this case, 40 bands in 200Hz intervals. The filter order was set to a constant value of 5 to ensure a balance between precision and computational efficiency.

The filter bank output $y(t)$ can be expressed as a sum of the outputs of individual bandpass filters:

$$y(t) = \sum_{i=1}^{N} g_i \cdot h_i(x(t)) \tag{1}$$

Here:

- N is the number of bands in the filter bank.
- g_i is the gain applied to the i-th band.
- $h_i(x(t))$ represents the output of the i-th bandpass filter applied to the input signal $x(t)$.

We calculate the RMSE difference between human and deepfake audio as a proportion so that it is normalised in reference to the null hypothesis. The null hypothesis would suggest that the amount of difference would be evenly distributed. The proportion of the RMSE difference within each frequency bin was calculated relative to the total difference across all bins. This can be expressed as:

$$\text{Proportional Difference} = \frac{\Delta(\text{Mean RMSE})}{\sum_{i=1}^{N} \Delta(\text{Mean RMSE}_i)} \tag{2}$$

The gain multiplier values were calculated by finding the normalised proportional amount of RMSE difference in each band and dividing by the sum difference as shown in Equation. 2. For Filter Bank 2 we squared the gain values to exaggerate the process, further decreasing similarities and increasing differences between the audio according to the analysis.

This method ensures that the gain values represent a proportionate contribution of each frequency band to the overall difference, making them relative to each other. This can be particularly important when dealing with signals of varying scales, as it ensures that the gain values reflect the proportional impact of each frequency band on the overall difference rather than being influenced by absolute magnitudes.

These proportional gain values were then multiplied by the number of bands. By scaling the normalized values by a factor of the number of bands, we effectively convert them into gain multipliers that reflect the relative difference between deepfake and authentic human voice in reference to the null hypothesis.

Additionally, we apply a cutoff filter that attenuates frequency bins that fall below the null hypothesis value. this ultimately resulted in a high-pass filter only allowing frequencies greater than 3201Hz.

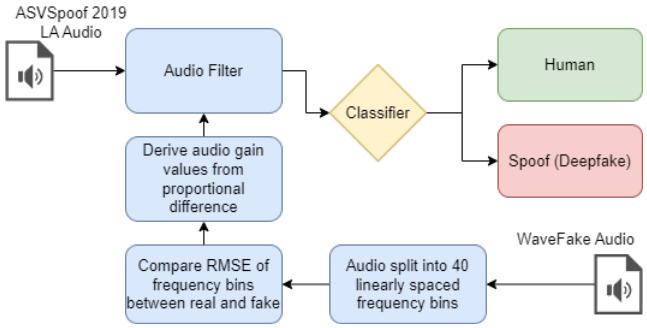

Fig. 3. System diagram with per-frequency band analysis of the WaveFake dataset.

III. Experimental Setup

Here we explain integrating our pre-processing step into a classification system can be seen in Figure. 3. Our primary goal was to test if our novel audio pre-processing technique could be used to improve a baseline model.

A. Datasets

To test whether this approach is applicable across different generative algorithms, we tested system performance by evaluating using the ASVSpoof 2019 LA dataset [14]. The ASVSpoof 2019 event separated the challenge into 2 more specific sub-challenges, the Logical Access (LA) and Physical Access (PA) challenges. This was to reflect the different nature of the two problems. We utilise the LA subset as the scope of our work is ASV countermeasures against deepfake speech. The ASVSpoof 2019 LA dataset focuses on testing against unseen generative methods, a general challenge in the area and a more in-the-wild scenario. All of the unseen generative methods in the ASVSpoof 2019 LA dataset were produced using neural vocoder methods generating audio waveforms in the mel-spectrum. All of the generated audio in this dataset was produced using the VCTK corpus [15] as base data. Many of these characteristics, such as using neural vocoders to generate waveforms from the mel-spectrum, are shared with the WaveFake dataset [16]. Only 6 of the generative methods for the ASVSpoof 2019 LA dataset occur in the training and development sets, the other 12 generative methods are unseen and are only present in the evaluation set to test generalisation ability. Table.I displays the makeup of the ASVSpoof 2019 LA dataset.

B. Baseline Model

The model we chose for the baseline model was the Squeeze-and-Excitation (SE) Residual Network (ResNet).

TABLE I
ASVSPOOF 2019 LA DATA PARTITION COUNTS.

Partition	#Bonafide	#Spoofed
Train	2,580	22,800
Dev.	2,548	22,296
Eval.	7,355	63,882

SEResNet is an extension of the traditional ResNet architecture incorporating a SE mechanism. The SE mechanism aims to improve the performance of Convolutional Neural Networks (CNN) by explicitly modelling the relationships between channels in the feature maps. We used the SEResNet50 variant of the model made open-source by Li et al. [17]. This model is one of the highest-performing CNN models when running as a baseline without the use of techniques such as feature fusion. Higher performance can be gained by the use of CQT or LFCC but our goal was to use a linearly scaled feature such as linearly scaled spectrograms to demonstrate the effectiveness of our novel filter technique. Features such as CQT use a logarithmic scale and diminish the information in the higher frequencies.

A consideration for choosing our baseline model was that the model should be a CNN-based model to allow for the use of Grad-Cam to produce insights. Grad-Cam visualisations localize important regions of an image input into a CNN model by mapping the various levels of activation in the model's final convolutional layer back to the input image space [18]. Grad-Cam allowed us to look into areas of focus in the spectrograms that the model learns throughout training. This allowed us to take a more white-box approach.

The model was trained using binary cross entropy as the loss function. The Adam optimizer was used with the parameters set to β_1 = 0.9, β_2 a= 0.98, and a weight decay of 10^{-9}. To manage the learning process and ease the model in, a learning rate scheduler was used for model warm-up, adjusting the learning rate dynamically during training to achieve the best possible performance. Initially, the learning rate steadily increases over the first 1000 warm-up steps, then gradually decreases proportionally to the inverse square root of the step number. The model underwent a 20-epoch training cycle, the model with the lowest EER on the development set was then chosen for evaluation.

C. Feature Representations

In our experimental setup, we used ordinally scaled magnitude spectrograms as the input feature for our model. The choice of spectrograms served a dual purpose: firstly, they provided a comprehensive and even representation of the frequency content, and secondly, their ordinally scaled nature controls for y-axis scaling. This deliberate choice enabled us to isolate and evaluate the unique contributions of our custom filter bank, as any observed improvements in model performance could be attributed specifically to the alterations introduced by the filter bank.

The spectrograms were extracted from the audio data using the STFT technique with parameters set as follows: *n_fft*=512, *hop_length*=160, *win_length*=400, and using the "hann" window function. These parameters were selected to strike a balance between temporal and frequency resolutions, ensuring a meaningful representation of the audio signal in the frequency domain.

IV. RESULTS

Our results show a decrease in both EER and Tandem Detection Cost Function (t-DCF), where lower is better as shown in Table.II. t-DCF is a more reliable predictor of performance when ASV and countermeasures are combined [19], it provides a more comprehensive accuracy measure that accounts for all four potential error scenarios arising from the interaction of the ASV system and the countermeasure system. On the other hand, EER is a more general metric for evaluating general authentication systems such as biometric systems. Our results suggest promise for this approach in the use of countermeasures to protect ASV systems as there is a negative trend in both EER and t-DCF. Our best result when compared to our baseline was using Filter Bank 2 where both EER and t-DCF were lower, providing an EER of 8.81 and a t-DCF of 0.1761. Other more complex features may benefit from this approach as the gain may reveal more discriminative factors in the higher frequencies. The results for our cutoff filter are in Table.II show very poor metrics indicating that while amplification of the higher frequencies can prove beneficial to classification performance ultimately they cannot be relied upon for even baseline-comparable performance.

TABLE II
RESULTS OF APPLYING THE BESPOKE FILTER BANKS. LINEAR FREQUENCY CEPSTRAL COEFFICIENTS (LFCC), CONSTANT Q CEPSTRAL COEFFICIENT (CQCC), GAUSSIAN MIXTURE MODELS (GMM)

Approach	EER	t-DCF
Our Baseline	10.334	0.1947
Filter Bank 1	9.923	0.1765
Filter Bank 2	8.810	0.1761
Cutoff Filter Bank	18.749	0.4239
ASVSpoof 2019 LA Baseline Systems [20]		
Baseline 1 (GMM CQCC)	9.57	0.2366
Baseline 2 (GMM LFCC)	8.09	0.2116

After the model was trained, we examined Grad-Cam activation heatmaps produced from the baseline SEResNet50 model with no filtering and after applying the filter to understand the performance, they can be seen in Figure. 4 and Figure. 5. Grad-Cam visualisations localize important regions of an image input in a CNN model, they allow us to see the amount of activation in certain regions of pixels as they pass through the CNN. Audio samples were randomly selected from the dataset to be visualised in a dip-sample. The Grad-Cam heatmaps show that there do exist areas in the mid to higher-frequency sections of the audio that are receiving a significant amount of focus from the baseline model, this is even more so after applying the filter. This shows that the filter has amplified regions of interest within the spectrogram, in doing so, the model does indeed use information in the higher frequencies as demonstrated by the increased levels of activation from Figure.

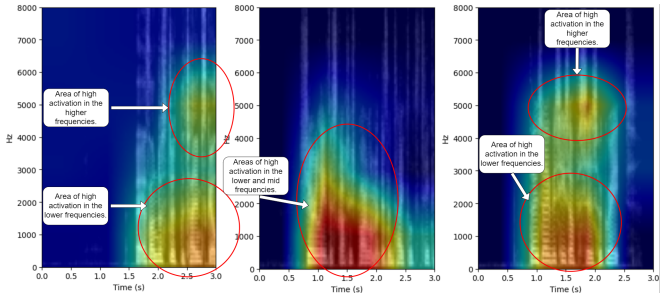

Fig. 4. Grad-Cam activation heatmaps of 3-second ASVSpoof 2019 LA audio samples, produced using the baseline SEResNet50 model and unfiltered audio.

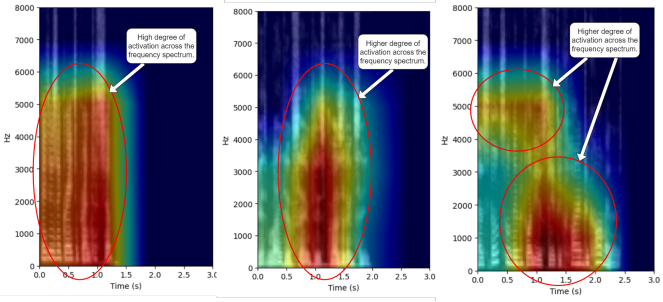

Fig. 5. Grad-Cam activation heatmaps of 3-second ASVSpoof 2019 LA audio samples, produced after filtering using the SEResNet50 model trained with Filter 2.

4 to Figure. 5. These portions would otherwise be diminished by using spectrograms where the y-axis is logarithmically or mel-scaled as a trade-off for focusing on the lower frequency portions of the spectrogram. Additionally, there appeared to be an area of less focus between many of the Grad-Cam visualisations between the 3kHz and 4kHz ranges shown in Figure 4, a similar pattern was also observed in the per-band analysis in Figure. 1. With these Grad-Cam visualisations, the performance of the cutoff filter starts to make more sense. While the higher frequencies can improve performance, as shown by our Filter Bank 1 and 2 results, cutting out the lower frequencies completely ruins performance. As the higher frequencies alone cannot be relied on for decent performance.

V. Conclusion

To conclude, we investigated whether there exist differences between human and deepfake audio that generalises across vocoders and how much of an improvement can be gained through exploiting these differences by using a novel filtering technique. Our analysis shows that there are generalisable differences between authentic human speech and deepfake-generated speech in the higher frequency ranges.

It is noteworthy that these frequency ranges, containing vital discriminative information, might receive limited spatial representation and reduced resolution in more conventional feature representations, such as mel and logarithmically scaled spectrograms. This information may shed light on new features for use in detecting deepfake voice as mel-spectrum that utilise differences in the higher frequency ranges in addition to using conventional features. From our results, these generalisable differences appear to show a trend of improvement by applying this novel technique at increasing strengths. Additionally, our Grad-Cam visualisations show that more of the spectrogram is being utilised after the audio has been filtered, specifically higher frequency regions.

In conclusion, this study represents a progression in understanding of differences between deepfake and genuine audio, particularly within the higher frequency range. Future avenues for investigation encompass the integration of this pre-processing step into SOTA models with a variety of deepfake data sets. This is imperative to effectively tackle the evolving challenges within audio forensics and synthetic speech detection.

References

[1] Enes Altuncu, Virginia N. L. Franqueira, and Shujun Li. *Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review*. 2022. arXiv: 2208.10913 [cs.CV].

[2] J. Damiani. *A Voice Deepfake Was Used to Scam a CEO Out of $243,000*. Accessed: 2021-10-14. 2019. URL: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

[3] Jaehyeon Kim, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5530–5540.

[4] Rodrigo Portugal and Benar Svaiter. "Weber-Fechner Law and the Optimality of the Logarithmic Scale". In: *Minds and Machines* 21 (Feb. 2011), pp. 73–81. DOI: 10.1007/s11023-010-9221-z.

[5] Akash Chintha et al. "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection". In: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 1024–1037. DOI: 10.1109/JSTSP.2020.2999185.

[6] Luigi Attorresi et al. "Combining automatic speaker verification and prosody analysis for synthetic speech detection". In: *arXiv preprint arXiv:2210.17222* (2022).

[7] Hemlata Tak et al. "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification". In: *arXiv preprint arXiv:2004.06422* (2020).

[8] Jichen Yang and Rohan Kumar Das. "Long-term high frequency features for synthetic speech detection". In: *Digital Signal Processing* 97 (2020), p. 102622. ISSN: 1051-2004. DOI: https://doi.org/10.1016/j.dsp.2019.102622. URL: https://www.sciencedirect.com/science/article/pii/S1051200419301769.

[9] Monisankha Pal, Dipjyoti Paul, and Goutam Saha. "Synthetic speech detection using fundamental frequency variation and spectral features". In: *Computer Speech & Language* 48 (2018), pp. 31–50. ISSN: 0885-2308. DOI: 10.1016/j.csl.2017.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0885230817301663.

[10] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. "A comparison of features for synthetic speech detection". In: (2015).

[11] Joel Frank and Lea Schönherr. *WaveFake: A Data Set to Facilitate Audio Deepfake Detection*. 2021. arXiv: 2111.02813 [cs.LG].

[12] Zhen Zeng et al. "Alignntts: Efficient Feed-Forward Text-to-Speech System Without Explicit Alignment". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6714–6718. DOI: 10.1109/ICASSP40776.2020.9054119.

[13] Keith Ito and Linda Johnson. "The lj speech dataset". In: (2017).

[14] Andreas Nautsch et al. "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.2 (2021), pp. 252–265. DOI: 10.1109/TBIOM.2021.3059479.

[15] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)". In: *University of Edinburgh. The Centre for Speech Technology Research (CSTR)* (2019).

[16] Xin Wang et al. "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech". In: *Computer Speech Language* 64 (2020), p. 101114. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2020.101114. URL: https://www.sciencedirect.com/science/article/pii/S0885230820300474.

[17] Xu Li et al. "Replay and Synthetic Speech Detection with Res2net Architecture". In: *arXiv preprint arXiv:2010.15006* (2020).

[18] Ramprasaath R Selvaraju et al. *Grad-CAM: Why did you say that?* 2017. arXiv: 1611.07450 [stat.ML].

[19] Tomi Kinnunen et al. *t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification*. 2019. arXiv: 1804.09618 [eess.AS].

[20] Massimiliano Todisco et al. "ASVspoof 2019: Future horizons in spoofed and fake audio detection". In: *arXiv preprint arXiv:1904.05441* (2019).