Interspeech 2024
1-5 September 2024, Kos, Greece

Unmasking Neural Codecs: Forensic Identification of AI-compressed Speech

Denise Moussa, Sandra Bergmann, Christian Riess

IT Security Infrastructures Lab, Friedrich-Alexander University of Erlangen-Nürnberg, Germany
{denise.moussa, christian.riess}@fau.de

Abstract

Compression traces are an important forensic cue to uncover the processing history and integrity of audio evidence. With continuous advances in the AI domain, efficient generative lossy neural codecs like Lyra-V2, EnCodec or Improved RVQGAN can compete with traditional speech and audio codecs. Their fundamentally different learning based approach compared to analytical lossy compression methods poses a new challenge for audio forensics. This calls for a closer examination of such techniques to prepare forensics for audio evidence processed by AI-based codecs. In this work, we thus want to take a first step towards robustly detecting traces of neural codecs in audio samples. We report that distinctive frequency artefacts enable for identifying neurally compressed audio and fingerprint specific AI-based codecs. We further analyse the robustness towards cross-dataset testing and noise, downsampling, and traditional compression post-processing.[1]

Index Terms: AI Audio Compression, Audio Forensics

1. Introduction

Lossy audio compression formats are ever-present in today's digital era and enable for efficient storage, sharing or real-time transmission of audio content. For audio forensics, lossy coding formats are both a challenge and can provide valuable clues. On the one hand, lossy compression can remove important cues and distort signals, such that forensic audio tools for, *e.g.*, speech splicing localisation [1] or the detection of synthetic speech [2] need to exhibit good robustness towards various coding formats. One the other hand, characteristic traces of lossy audio codecs can also be forensically exploited, for example to uncover manipulated parts of audio signals [3].

Up to now, forensic research focuses on traditional lossy coding formats like MP3, Vorbis or Opus that use established techniques like transform and perceptual coding [4] to effectively compress signals via quantisation while modelling the human hearing system. However, recent advances in the domain of neural codecs provide highly competitive and fundamentally different learning-based models that perform successfully at extremely low bitrates [5–14]. This development requires forensic tools to be ready to handle this new type of lossy compression formats. Recently, forensic examination of AI-based compression and the detection of it has moved into focus for image data [15–17]. However, forensic investigation of neural audio compression has to the best of our knowledge not yet been examined.

In this work, we therefore take a first step to analyse the robust detection and identification of neural audio codecs. We conduct our study on the forensically important case of speech signals and select three practically relevant open-source neural codecs. This includes Google's© Lyra-V2 [8] speech codec[2] that incorporates a further development of the SoundStream [12] model, Meta's© EnCodec [13] network[3] and Descript Inc.'s© very recently proposed Improved RVQGAN [14][4]. Our detailed contributions thus are:

- To identify variations in the frequency representation of neural codecs that lead to distinct peak artefacts, especially in the high frequency domain.
- To demonstrate the forensic exploitability of the artefacts for lossy neural compression detection and neural codec fingerprinting in a cross-dataset setting.
- To demonstrate good robustness in increasingly difficult testing conditions including unseen codecs and post-processing, particularly when using hand-crafted features.

2. Related Work

The detection of artefacts from traditional lossy coding formats has a long history in the research field of audio forensics to uncover the processing history of a signal. Multiple compression runs from re-saving operations or inconsistent coding traces within a file can hereby be used to detect or localise potential manipulation operations [3]. Codec-independent single compression detection has been targeted using analytical time-frequency features [18] or deep convolutional neural network (CNN) features [19]. Additionally, stacked autoencoders have been explored to extend the task up to fourth-time compression detection [20]. A large number of works specifically target the analysis of the MP3 format due to its great popularity and practical relevance. By example, statistical scale factor features have been proposed for double MP3 compression detection [21] with the same bitrate. Yan *et al*. [22] additionally exploit Huffman table indices to discriminate single, double and triple MP3 compressed audio. Recently, the use of Transformers [23] has been proposed to localise multiple MP3 compressed sections within some signal to uncover splicing manipulations [3].

Works on synthetic speech detection similarly often rely on neural vocoder artefacts. Here, for example, Pons *et al*. [24] report upsampling artefacts from transposed convolutions in the decoder of a neural audio synthesiser. Morrison *et al*. [25] show that many GAN vocoder architectures suffer from pitch and periodicity errors. Also, learned deep features exhibit good empirical performance to distinguish real human speech and synthetic speech from seen and unseen vocoders [26].

[1] code: https://faui1-gitlab.cs.fau.de/mmsec/forensic-identification-of-ai-compressed-speech

[2] https://github.com/google/lyra

[3] https://github.com/facebookresearch/encodec

[4] https://github.com/descriptinc/descript-audio-codec

10.21437/Interspeech.2024-1652

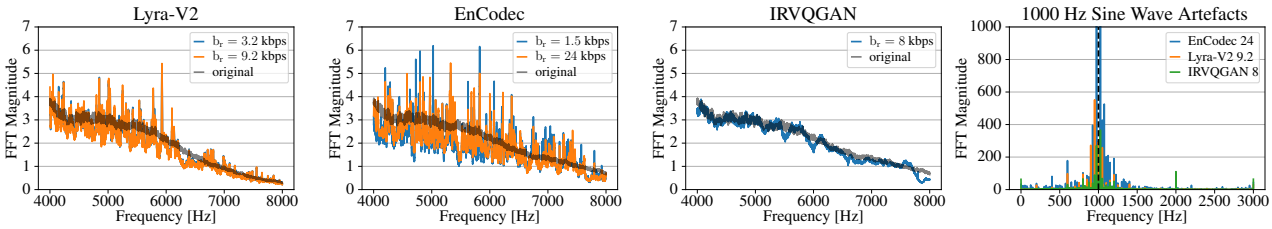


Figure 1: *Left: averaged FFT for* $4-8$ *kHz of* 500 *LibriSpeech samples per neural codec. Right: artefacts of a* 1 *kHz signal.*

To the best of our knowledge, the identification and detection of neural audio codecs has not yet been explored. Still, the importance of forensic treatment of neural compression has been recently recognised in the image domain [15–17]. Berthet *et al*. [15, 16] exemplarily show the inability of JPEG compression detectors to generalise to AI-compressed images and highlight their impact on forensic tools, and Bergmann *et al*. [17, 27] provide a first analysis on neural image codec traces.

3. Methods

We summarise neural audio codecs in Sec. 3.1, describe our datasets in Sec. 3.2, and discuss characteristic compression artefacts of neural codecs in the frequency domain in Sec. 3.3.

3.1. Generative Neural Audio Compression

Contrary to traditional audio codecs that use analytical models, fully neural audio codecs employ a purely data-driven approach to learn efficiently compressed representations from original signals. In the following, we outline the technical principles behind the three end-to-end trainable audio codecs selected for this study [12–14], and highlight their main differences.

Principles of Neural Audio Codecs Neural codecs typically employ an encoder $\mathrm{E}(\cdot)$, a vector quantiser $\mathrm{V_q}(\cdot)$ and a decoder $\mathrm{D}(\cdot)$ to produce a lossy version $\hat{\mathbf{s}} \in \mathbb{R}^T$ of the input signal $\mathbf{s} \in \mathbb{R}^T$ of length T as $\hat{\mathbf{s}} = \mathrm{D}(\mathrm{V_q}(\mathrm{E}(\mathbf{s})))$ [8, 12–14]. In detail, an input audio signal $\mathbf{s} \in \mathbb{R}^T$ is processed by the encoder $\mathrm{E}(\cdot)$ which outputs a series of latent representations $\mathbf{H} = [\mathbf{h}_0, \mathbf{h}_1, \cdots, \mathbf{h}_N]$, where $h_n \in \mathbb{R}^D$ with latent size D. A vector quantiser $\mathrm{V_q}(\cdot)$ then compresses $\mathbf{H}$ to some target bitrate, usually denoted in kilobits per second (kbps). In detail, our selected models employ residual vector quantisation (RVQ) adapted from the Vector Quantised-Variational AutoEncoder (VQ-VAE) model [28, 29]. The decoder reconstructs the time-domain signal $\hat{\mathbf{s}} \in \mathbb{R}^T$ from the quantised encoder output $\hat{\mathbf{H}}$. During training, a mixture of adversarial losses and signal reconstruction losses is combined to yield compressed signal representations with high fidelity [12–14].

Neural Codec Models We investigate three officially released, pre-trained neural codecs, namely Lyra-V2 [8, 12], EnCodec [13] and Improved RVQGAN [14], further referred to as IRVQGAN.[2,3,4] SoundStream, which introduced the efficient RVQ in an end-to-end trainable model [12], is used in improved form in Lyra-V2. EnCodec builds up on SoundStream and proposes sequential modelling, an improved adversarial loss and loss balancing [13]. IRVQGAN uses similar ideas, but focuses on high-quality signal reconstruction at the expense of longer computation times and five times more model parameters (*c.f*. Tab. 1). The authors employ periodic activation functions, mitigate codebook collapse [30], and apply state-of-the-art vocoder training strategies to further optimise reconstruction fidelity.

For our study, we select provided standard versions of the codecs. Lyra-V2 uses a sampling rate of 16 kHz, EnCodec compresses speech at 24 kHz, and we equally choose IRVQGAN on 24 kHz. All models output audio with a bit depth of 16 bit. We analyse all supported bitrates b_r, *i.e*., $b_r \in \{3.2, 6, 9.2\}$ kbps for Lyra-V2, $b_r \in \{1.5, 3, 6, 12, 24\}$ kbps for EnCodec, and $b_r = 8$ kbps for IRVQGAN.

3.2. Dataset Selection

Our analysis is done on speech samples that none of the codecs has trained on [12–14]. We further test on sets of different signal characteristics, since robustness is an important factor in forensics. Thus, we select samples both from the 16 kHz, 16 bit, LibriSpeech [31] test-clean split which is constructed from read audiobook data and the 48 kHz, 16 bit TSP [32] database that (contrary to LibriSpeech) provides noiseless, anechoic speech signals. We sample balanced pools of 2 s speech snippets from female and male speakers. In total, two training pools of 1000 samples are constructed each from the LibriSpeech and TSP database, where a random $90/10$ train/validation split is applied for experiments. For testing, we use the remaining 184 samples from TSP and a larger distinct pool of 500 LibriSpeech samples. More details on data preparation and post-processing are described in the respective experiments in Sec. 4.

3.3. Artefacts of Neural Codecs

We examine the fast Fourier transform (FFT) of AI-compressed audio for artefacts. Figure 1 shows averaged frequency spectra from speech samples from 4 to 8 kHz for the lowest and highest bitrate setting of Lyra-V2 and EnCodec and the single available bitrate for IRVQGAN. The spectrum of the original version and its compressed counterparts are shown in black and in colour. Evidently, the neural codecs make notable frequency reconstruction errors that especially occur in the plotted high frequency range. The artefacts are comparable for the two Lyra-V2 bitrate settings, while EnCodec's error increases with much stronger quantisation. IRVQGAN best represents the frequencies, but deviations from the original are still noticeable.

The noisy FFT spectrum stems from the tendency of neural codecs to distribute the energy of specific frequencies across the whole frequency spectrum. Figure 1 (right) shows a specific example for a compressed single tone signal of 1 kHz with 2 s duration. All neural codecs put the highest energy on the 1 kHz frequency, with coefficient magnitudes of about 20,000 for EnCodec and IRVQGAN, and about 6,000 for Lyra-V2 (clipped in Fig. 1 for better visualisation). However, a significant amount of energy is distributed in close proximity of the original frequency and artefacts are still visible far off the target frequencies.

In addition, to better see the reconstruction ability for specific frequencies, each neural codec is fed with 2 s signals of single frequency $f \in [20, 8000]$ Hz in 50 Hz steps. We report the relative amount of reconstructed energy E_r versus original

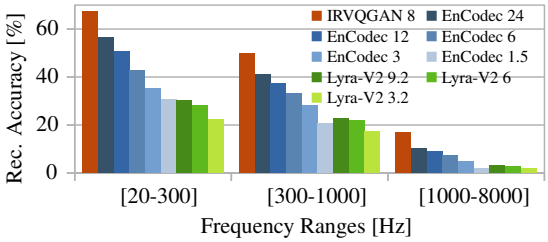


Figure 2: *Reconstruction accuracy of neural codecs.*

Table 1: *Accuracy (avg* [%] ± *std. dev.* [%]*) of* 5 *Logistic Regression training runs for detecting AI-compressed speech from a specific neural codec on LibriSpeech and TSP samples.*

Model	Param.	Lib.→Lib.	Lib.→TSP	TSP→Lib.	TSP→TSP
Lyra-V2	12 M	96.41±0.8	96.17±0.6	94.53±0.8	97.40±1.0
EnCodec	15 M	99.76±0.2	98.77±0.5	85.28±2.1	100.00±0.0
IRVQGAN	75 M	98.70±0.0	85.87±0.0	79.40±0.0	100.00±0.0

energy E_f as $\text{acc}_r = \frac{E_r}{E_f}$. Figure 2 shows the results, binned into three frequency ranges. IRVQGAN achieves the highest reconstruction ratios, followed by EnCodec and Lyra-V2. Lower frequencies until 300 Hz are reconstructed with overall higher accuracy, which also includes frequencies of human speech. The higher the frequencies, the lower is the reconstruction fidelity. This may stem from the fact that neural codecs are mainly trained on speech and music [12–14], and hence more strongly constrained in lower frequency bands.

In this work, we focus on a practical exploitation of such irregularities. We investigate the detectability of the codecs using the FFT of speech signals. To enhance strong amplitude changes in the signal we process each sample $\mathbf{s}$ with a high-pass filter, *i.e.*, the kernel $\mathbf{k}_L = [-1, 2, -1]$, compute the magnitude of the FFT and convert the output to logarithmic scale, yielding $\text{FFT}^\star(\mathbf{s}) = \log(|\,\text{FFT}(\mathbf{s} * \mathbf{k}_L)|)$. In our studies, this enhancement leads to strongly improved detection results.

4. Detecting Traces of Neural Codecs

The experiments show the robust detectability and identification of neural codecs with the discussed $\text{FFT}^\star$ features under forensically challenging settings. We use a Logistic Regression classifier and compare to deep features from several neural network (NN) classifiers. Reported metrics are the mean and standard deviation (std. dev.) of 5 training runs with different seeds.

4.1. Intra-Codec Detectability

This experiment investigates the discriminative power of the $\text{FFT}^\star$ features for distinguishing uncompressed and AI-compressed speech signals of a specific neural codec. All data pools are resampled to the supported sampling rate of the respective codec model. Both train/test pools (LibriSpeech and TSP) are compressed by each codec for all supported bitrates $\text{b}_r \in \mathcal{B}$ (*c.f.* Sec. 3.1) to a total of $2 \cdot 9$ AI-compressed sets. Each set is paired with its uncompressed, resampled counterpart to train a binary Logistic Regression classifier with a random 90/10 train/validation split. We test within and across datasets. Lyra-V2 and EnCodec support multiple bitrates and are hence evaluated within and across bitrates. Results from all $|\mathcal{B}|^2$ bitrate combinations are averaged to yield compact tables.

Table 2: *Accuracy (avg* [%] ± *std. dev.* [%]*) of* 5 *Logistic Regression training runs on IRVQGAN samples for detecting AI-compressed speech signals from unseen neural codecs.*

Model	Lib.→Lib.	Lib.→TSP	TSP→Lib.	TSP→TSP
Lyra-V2	50.43±0.7	85.96±0.8	61.03±0.3	98.55±0.1
EnCodec	86.22±1.7	94.40±0.9	75.20±2.8	99.68±0.1

Table 1 shows the results. All models are robustly detected, averaging to $96.13\% > 95.95\% > 90.99\%$ for Lyra-V2, EnCodec and IRVQGAN. As expected, models with stronger frequency artefacts are detected with higher accuracy (*c.f.* Sec. 3.3), and generalisation from anechoic TSP to LibriSpeech samples is most challenging. The std. devs. across runs and bitrate changes are quite low. IRVQGAN achieves std. devs. of 0% due to its fixed bitrate and of course due to the stable convergence of Logistic Regression. Lyra-V2 and EnCodec also generalise well with low std. devs. across bitrates since both exhibit notable artefacts for all bitrates (*c.f.* Sec. 3.3).

4.2. Generalisation to Unseen Neural Codecs

We further analyse the capability to generalise to the detection across neural codecs, i.e., to train on one codec and to try to detect compression from another codec. The remaining experimental setup (cross-dataset, cross-bitrate, and train/test composition) is identical to the previous experiment.

The results show that generalisation is partly challenging. After training on 16 kHz Lyra-V2 compressed samples, the detection performance on the other codecs drops to almost guessing with an accuracy of 53.20%. Also, training on 24 kHz EnCodec data may achieve a cross-codec performance of 81.89% for IRVQGAN but only 58.90% for Lyra-V2. Overall, training on the 24 kHz IRVQGAN with its more subtle artefacts generalises better to Lyra-V2 and EnCodec, which is shown in Tab. 2. Here, EnCodec and Lyra-V2 are detected with an average accuracy of 88.87% and 73.99%. Hence, cross-codec generalisation tends to be easier when training on data with less pronounced artefacts than present in test samples.

4.3. Fingerprints of Neural Codecs

A common forensic question is to identify the specific codec that has been used for compression. We perform a closed-set 4-class classification to distinguish uncompressed samples from Lyra-V2, EnCodec, and IRVQGAN data. The bitrates for Lyra-V2 and EnCodec are set to 9.2 kbps and 24 kbps for best-quality compression with smallest artefacts. We construct two train/test sets of 4K/2K and 4K/736 samples from LibriSpeech and TSP, where each set includes 4 versions of each sample, *i.e.*, one for each class. All signals are downsampled to the lowest common sampling rate of 16 kHz and Logistic Regression is trained as one-versus-rest classifier on both train sets.

The first row of Tab. 3 shows the F1-Scores for each class for the best (LibriSpeech/LibriSpeech) and worst (LibriSpeech/TSP) train/test set combinations. In most cases, codecs are robustly identified with 0% std. dev., however, difficult cases occur in the cross-set setting, with accuracies of 63.64% on uncompressed samples and of 56.90% on IRVQGAN samples. The confusion matrices in Fig. 3 (left) further show that these two classes get mixed up upon generalisation to TSP. This is again in agreement with the fact that IRVQGAN leaves the weakest artefacts (*c.f.* Sec. 3.3).

Table 3: *F1-Score (avg [%] ± std. dev. [%]) of 5 runs for classifying uncompressed, Lyra-V2, EnCodec and IRVQGAN samples.*

		Uncompressed		Lyra-V2 (9.2 kbps)		EnCodec (24 kbps)		IRVQGAN (8 kbps)	
Classifier	Param.	Lib.→Lib.	Lib.→TSP	Lib.→Lib.	Lib.→TSP	Lib.→Lib.	Lib.→TSP	Lib.→Lib.	Lib.→TSP
Logistic Regression	16 K	88.02±0.0	63.64±0.0	95.01±0.0	90.08±0.0	99.20±0.0	99.46±0.0	92.83±0.0	56.90±0.0
ResNet-18	11 M	99.90±0.1	66.59±0.2	99.98±0.0	98.33±0.5	99.94±0.1	99.67±0.4	99.98±0.0	00.00±0.0
EffNet-B0	4 M	99.02±0.5	65.68±0.5	99.54±0.4	98.10±1.0	99.88±0.1	99.45±0.6	99.56±0.3	01.05±2.1
RegNetY-400mf	4 M	98.32±1.4	65.18±0.8	99.24±0.6	98.44±0.4	99.47±0.6	98.44±1.1	99.64±0.2	00.00±0.0

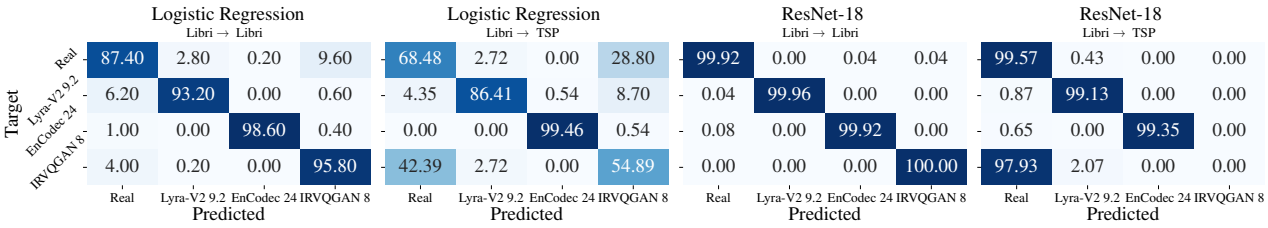


Figure 3: *Averaged, normalised confusion matrices over 5 training runs for Logistic Regression and the best NN classifier in %.*

4.3.1. Comparison to Deep Features

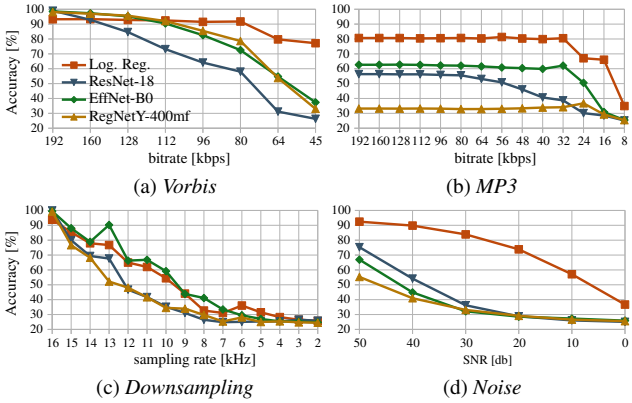
We compare to deep features extracted by the standard classifiers ResNet-18 [33], EffNet-B0 [34] and RegNetY-400mf [35] on the n-class task. For this, a linear layer of output size 4 is added to each model's feature extractor. To account for the NNs' greater need for training data, we scale up the LibriSpeech pool with the remaining samples from the same source split, yielding about 7 K samples. The distinct test pool remains as is. The final train/test sets are constructed as in Sec. 4.3, where standard short-time Fourier transform (STFT) representations with the same enhancements as for the FFT* features are computed with window length $w = 800$ and hop size $\frac{w}{2}$ from each signal. All models converge within 100 epochs, with a batch size of 128 and the Adam optimiser with learning rate $1e^{-3}$.

Table 3 shows the results. In many cases, the NN classifiers outperform Logistic Regression, however they are more prone to overfitting on specific data characteristics. Thus, when testing on the TSP set, all neural classifiers identify the low-artefact IRVQGAN samples as uncompressed signals with an F1-Score around 0%. Figure 3 (right) further shows the excellent intra-set but poor generalisation ability of the best performing ResNet-18 [33]. Therefore, further work has to be invested to design powerful, robust neural detectors that can compete with the low-cost, explainable Logistic Regression classifier on handcrafted features.

4.3.2. Feature Robustness to Post-Processing

In forensic use cases, audio signals may stem from uncontrolled sources and might be subject to post-processing that weakens forensic traces. Fig. 4 shows the results for the 4-class experiment on the LibriSpeech test set with additional post-processing.

Multi-Compression Audio content is routinely recompressed upon sharing over the internet. Recompression with traditional codecs tends to remove high-frequencies due to perceptual coding [4], which can weaken the traces of neural codecs. Fig. 4a and Fig. 4b show the detection results after recompression with Vorbis and MP3 compression. Vorbis achieves higher quality signals than MP3 at equal bitrates, which reflects in the classifiers' performance. Logistic Regression exhibits higher robustness than NNs, with a noteworthy accuracy above 80% for MP3 with $b_r \geq 32$.



(a) *Vorbis* (b) *MP3*

(c) *Downsampling* (d) *Noise*

Figure 4: *Accuracy (avg [%]) of 5 runs on the 4-class task.*

Downsampling Signal downsampling is critical, since all information above the Nyquist Frequency gets removed. Thus, all classifiers are strongly impacted as shown in Fig. 4c. Logistic Regression still performs best together with EffNet-B0 [34].

Noise General signal perturbations can be simulated by additive Gaussian white noise. Fig. 4d shows that Logistic Regression gently degrades with decreasing signal-to-noise ratio (SNR), while NNs are very sensitive to this type of degradation.

5. Conclusion

In this work, we show that distinct frequency artefacts can be used to detect AI-compressed speech and identify specific neural codec architectures. This is especially the case for the real-time speech codecs Lyra-V2 and EnCodec, while the higher quality and computationally more expensive IRVQGAN is harder to detect, especially in out-of-distribution classification scenarios. Our analytic and explainable FFT* features yield good results, while standard NNs show overfitting issues. Yet, we hope that this first study motivates further research for the exploration of more powerful and robust NN classifiers.

6. Acknowledgements

We kindly thank our colleagues from the Federal Criminal Police Office of Germany for their support in this research project.

7. References

[1] D. Moussa, G. Hirsch, S. Wankerl, and C. Riess, "Point to the Hidden: Exposing Speech Audio Splicing via Signal Pointer Nets," in *Proc. of Interspeech*. ISCA, 2023, pp. 5057–5061.

[2] L. Cuccovillo, M. Gerhardt, and P. Aichroth, "Audio Spectrogram Transformer for Synthetic Speech Detection via Speech Formant Analysis," in *IEEE International Workshop on Information Forensics and Security*. IEEE, 2023, pp. 1–6.

[3] Z. Xiang, P. Bestagini, S. Tubaro, and E. J. Delp, "Forensic Analysis and Localization of Multiply Compressed MP3 Audio Using Transformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 2929–2933.

[4] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[5] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet Based Low Rate Speech Coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 676–680.

[6] C. Gârbacea, A. van den Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low Bit-Rate Speech Coding with VQ-VAE and a WaveNet Decoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 735–739.

[7] F. S. Lim, W. B. Kleijn, M. Chinen, and J. Skoglund, "Robust Low Rate Speech Coding Based on Cloned Networks and Wavenet," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 6769–6773.

[8] W. B. Kleijn, A. Storus, M. Chinen, T. Denton, F. S. Lim, A. Luebs, J. Skoglund, and H. Yeh, "Generative Speech Coding with Predictive Variance Regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 6478–6482.

[9] S. Li, H. H. Mao, and J. McAuley, "Variable Bitrate Discrete Neural Representations via Causal Self-Attention," in *2nd Preregistration workshop*, 2021.

[10] T. Jayashankar, T. Koehler, K. Kalgaonkar, Z. Xiu, J. Wu, J. Lin, P. Agrawal, and Q. He, "Architecture for Variable Bitrate Neural Speech Codec with Configurable Computation Complexity," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 861–865.

[11] X. Jiang, X. Peng, C. Zheng, H. Xue, Y. Zhang, and Y. Lu, "End-to-End Neural Speech Coding for Real-Time Communications," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 866–870.

[12] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An End-to-End Neural Audio Codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[13] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Transactions on Machine Learning Research*, 2023.

[14] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] A. Berthet and J.-L. Dugelay, "AI-based Compression: A New Unintended Counter Attack on JPEG-Related Image Forensic Detectors?" in *IEEE International Conference on Image Processing*. IEEE, 2022, pp. 3426–3430.

[16] A. Berthet, C. Galdi, and J.-L. Dugelay, "On the Impact of AI-Based Compression on Deep Learning-Based Source Social Network Identification," in *IEEE 25th International Workshop on Multimedia Signal Processing*. IEEE, 2023, pp. 1–6.

[17] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, "Forensic Analysis of AI-compression Traces in Spatial and Frequency Domain," *Pattern Recognition Letters*, 2024.

[18] B. Kim and Z. Rafii, "Lossy Audio Compression Identification," in *26th European Signal Processing Conference*. IEEE, 2018, pp. 2459–2463.

[19] R. Hennequin, J. Royo-Letelier, and M. Moussallam, "Codec Independent Lossy Audio Compression Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 726–730.

[20] D. Luo, W. Cheng, H. Yuan, W. Luo, and Z. Liu, "Compression Detection of Audio Waveforms Based on Stacked Autoencoders," in *Artificial Intelligence and Security: 6th International Conference*. Springer, 2020, pp. 393–404.

[21] P. Ma, R. Wang, D. Yan, and C. Jin, "Detecting Double-Compressed MP3 with the Same Bit-rate," *J. Softw.*, vol. 9, no. 10, pp. 2522–2527, 2014.

[22] D. Yan, R. Wang, J. Zhou, C. Jin, and Z. Wang, "Compression History Detection for MP3 Audio," *KSII Transactions on Internet & Information Systems*, vol. 12, no. 2, 2018.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling Artifacts in Neural Audio Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 3005–3009.

[25] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked Autoregressive GAN for Conditional Waveform Synthesis," in *International Conference on Learning Representations*, 2022.

[26] C. Sun, S. Jia, S. Hou, and S. Lyu, "AI-Synthesized Voice Detection Using Neural Vocoder Artifacts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 904–912.

[27] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, "Frequency-Domain Analysis of Traces for the Detection of AI-based Compression," in *11th International Workshop on Biometrics and Forensics*. IEEE, 2023, pp. 1–6.

[28] A. Van Den Oord, O. Vinyals *et al.*, "Neural Discrete Representation Learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[30] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized Image Modeling with Improved VQGAN," in *International Conference on Learning Representations*, 2022.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 5206–5210.

[32] P. Kabal, "TSP Speech Database," *McGill University, Database Version*, 2002.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[34] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[35] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing Network Design Spaces," in *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2020, pp. 10 428–10 436.