IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Speech Watermarking removal by DNN-based Speech Enhancement Attacks

Álvaro López-López, Eros Rosello, Angel M. Gomez

Dept. Signal Processing, Telematics and Communications - CITIC, Universidad de Granada, Spain
alvaro192@correo.ugr.es, erosrosello@ugr.es, amgg@ugr.es

Abstract

Audio watermarking allows for embedding a bitstream in an audio file while ensuring the introduced alteration remains imperceptible. Although its primary application has been in copyright protection, it has recently been proposed as a proactive and robust method to detect synthetic speech. By marking artificial speech, voice deepfakes can be easily identified, preventing their misuse regardless of the naturalness or realism achieved by sophisticated generative speech models. However, watermarks could be subjected to tampering and removal attacks, which aim to disguise deepfakes as genuine speech. Recently, deep learning approaches have been applied to watermarking, resulting in highly reliable and resilient speech watermarkers. Nonetheless, the very same approaches can be followed by attackers. In this paper, we propose a novel DNN-based removal attack and evaluate its effectiveness against both classical and deep learning-based watermarking methods. This attack leverages a well-known speech enhancement architecture, the DCCRN model, and, as we demonstrate, it achieves remarkable success in removing watermarks while maintaining excellent speech quality and intelligibility.

Index Terms: Watermarking, Tampering and removal, Speech enhancement, Deep learning

1. Introduction

Generative speech models have reached a point where they can synthesize voices that are almost indistinguishable from real ones [1–3]. Although these technologies have clear legitimate applications [4, 5], they can also be used for malicious purposes, introducing new security risks which have become increasingly apparent in the recent years, with the rise of deepfakes, voice hoaxes, scam calls, and frauds [6, 7]. The threat escalates even further when synthetic voice is used to spoof automatic speaker verification (ASV) and biometric systems, and hence the research community has made significant efforts to provide countermeasures [8, 9]. However, the primary forensic method for detecting synthesized audio involves training binary classifiers to distinguish between natural and artificial speech, as in [10–12]. Nonetheless, these methods are likely to become ineffective as the gap between synthesized and genuine speech narrows.

For this reason, audio watermarking has emerged as a promising alternative: by subtly embedding marks in the output signal of the synthesizer, synthesized speech can be easily identified. As a disadvantage, this proactive approach requires the collaboration of the synthesis system, which could be enforced [13] for high-quality and realistic ones, before making them available to the public.

Traditional audio watermarking methods rely on embedding watermarks in either time or frequency domain [14–17], often incorporating domain-specific features to design the watermark and its corresponding decoding function. These methods heavily depend on expert knowledge and empirical rules, resulting in low encoding capacity and limited resilience to attacks [18]. To address these limitations, deep learning approaches have recently been applied to audio watermarking, offering enhanced robustness and efficiency [19]. Particularly, two DNN-based systems have been proposed specifically for speech watermarking [20, 21], showing a remarkable resilience to a broad range of tampering attacks.

In this paper we investigate these attacks, where an impostor attempts to remove the watermark in order to disguise synthetic speech as genuine. In particular, we propose a novel approach based on a deep learning architecture. DNN-based attacks have shown very promising results in the image processing field [22, 23], but to the best of our knowledge, they have not yet been tested against speech watermarking. Thus, our approach leverages a well known speech enhancement (SE) architecture which is specifically trained to remove audio watermarks from speech signals. As we show, this architecture is very effective at damaging the watermarks, especially those introduced by DNN-based methods, while keeping an outstanding speech quality.

The rest of this paper is organized as follows. In Section 2, we describe the materials and methods used in our experiments. This includes the speech database, the watermarking techniques, the attacking methods —particularly the proposed DNN-based enhancement technique— and the performance metrics considered. In Section 3, we present the results obtained and discuss the effectiveness of the tested methods. Finally, conclusions are summarized in Section 4.

2. Materials and methods

2.1. Speech database

Watermarking techniques and DNN-based attacking methods have been tested and trained on an in-house datasets derived from the TIMIT database [24]. This database is comprised of approximately five hours of English speech with broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. Our customized datasets consist of utterances randomly selected from the same speaker, downsampled to 16 kHz and concatenated so that the resulting signal's length is between 7 and 11 seconds. Such a length has been chosen to facilitate the watermarking procedure as well as the speech evaluation (according to ITU recommendation [25]). Following this procedure, three different datasets have been created for DNN training, validation, and testing, containing 1656, 339, and 361 utterances, respectively.

10.21437/IberSPEECH.2024-4

To ensure independence, no speakers are repeated across these datasets. Additionally, speaker gender is balanced within each dataset.

2.2. Watermarking techniques

In this paper we consider a total of 6 different watermarking (WM) techniques ranging from classical methods to state-of-the-art DNN-based approaches. These are briefly described below.

Each WM method is applied to the aforementioned training, validation and testing datasets. To do so, a random bit sequence is generated, associated to and embedded in each utterance. The length of this sequence is in accordance with the capacity of the WM technique and the length of the utterance to be marked.

2.2.1. DCT-based watermarking

An efficient and robust watermarking technique fully based on the Discrete Cosine Transform (DCT) is proposed in [14]. This method operates on a frame-by-frame basis, splitting each frame into two sections: a small section for smooth transitions and a larger section converted into DCT coefficients. Since most energy is concentrated below 1 kHz, only the first DCT coefficients are used, divided into three bands. In our implementation, only the first band is utilized.

The watermark remains imperceptible as long as the energy distortion within the frame falls below the masking threshold. To achieve this, the DCT coefficients are randomly divided into two groups: one for embedding information and the other for absorbing energy fluctuations. During the embedding stage, a perceptual quantization index modulation (QIM) rule [26] is applied to minimize quality degradation. As a result, the embedded watermark is imperceptible to human ears and robust against common digital signal processing attacks.

2.2.2. Patchwork

The patchwork scheme [15] allows a multilayer embedding which is able to embed watermarks to a host media signal repeatedly in an overlaying manner, without interference between layers. For simplicity and due to the fact that adding multiple layers of watermark bits inevitably reduces the perceptual quality, in this work, only the first layer is used.

This method also uses a DCT, but computed across the entire signal. Only the DCT coefficients within the frequency range [3000, 7000] Hz are selected and partitioned into several so-called DCT segments. To ensure imperceptibility, the method requires that the average magnitude of all these segments to be similar. Since energy is concentrated in the low frequency components, lower and higher DCT coefficients are mixed and reordered in a specific manner prior this partitioning. Each bit is then embedded into two rearranged consecutive segments, known as a fragment pair, by modifying the relative average magnitudes of each one.

As the whole signal is used to compute DCT segments, this method is particularly resilient, but at the cost of a lower capacity.

2.2.3. FSVC watermarking

Another method which makes use of the DCT, but in conjunction with a Singular Value Decomposition (SVD), is the Frequency Singular Value Coefficient (FSVC) watermarking scheme [16]. In this method, the audio signal is segmented into many frames as the number of data bits to be embedded. Each frame is then split into two segments of equal length and a DCT is applied to each of them.

After selecting the DCT coefficients within the range from 1.09 to 1.45 kHz frequencies, a SVD is applied to them. The FSVC coefficient is determined as the ratio between the pair of singular value coefficients from both segments. The embedding process is performed by manipulating this ratio. By following this procedure the watermark can resist desynchronization attacks in addition to common digital signal processing attacks.

2.2.4. Norm-space watermarking

As FSVC, the Norm-space watermarking method [17] also segments the audio signal into multiple frames, corresponding to the number of data bits to be embedded. For each frame, a Discrete Wavelet Transform (DWT), based on Haar wavelets, is used to extract a sequence of approximation coefficients. A DCT transform is then applied, producing a vector V, which is then subsampled into V_1 and V_2 by selecting even and odd indices, respectively.

The L2-norm of each vector is computed and averaged to obtain a global norm value for the frame ($norm = (||V_1||_2 + ||V_2||_2)/2$). During the embedding process, the norms of V_1 and V_2 are shifted by $\pm\Delta$ from this norm (each in opposite directions), where the sign indicates whether the bit to be encoded is a one or a zero.

The authors of this method claim that it makes an excellent tradeoff between security, capacity, imperceptibility and robustness against signal processing attacks. However, we have observed that it is particularly sensitive to fully silent frames, where norms approach zero and watermarks are easily lost.

2.2.5. WavMark

WavMark [20] is a state-of-the-art DNN-based watermarking scheme characterized by an invertible encoder/decoder architecture, a shift module, and an attack simulator module.

Encoding and decoding processes are carried out by means of an Invertible Neural Network (INN) [27]. This kind of network operates bidirectionally through invertible transformations, so a forward pass maps complex data distributions into a simple latent distribution. Conversely, a backward pass (where inputs are computed from outputs) regenerates the data distribution from the latent space. Both original audio signal and bit sequence are used as inputs, while the watermarked signal is considered as the output of the network.

To ensure a successful decoding even when there are displacements in the decoding position, the embedding process constructs the encoded message by combining pattern and payload bits. A brute force detection is then applied, where the module shifts the decoding window along the temporal direction by a random time length. The pattern bits are used for checking the validity of the decoded outputs. In addition, an attack module is applied on the watermarked signal during DNN training, simulating ten common signal processing attacks, such as random noise, sample suppression, low-pass filtering, resampling, or time stretching.

A curriculum learning strategy is adopted to force the model to progressively acquire encoding capabilities through three distinct stages. In the first one, the attack simulator is excluded and weak constraints are imposed on perceptibility. In the second, the attack simulator is introduced. Finally, in the third one, strong perceptual constraints are enforced to ensure that watermarks remain imperceptible.

2.2.6. AudioSeal

AudioSeal is a recent watermarking method that, like WavMark, follows a DNN approach [21]. Two interconnected models, a generator and a detector —both derived from a neural audio compression architecture [28]— are jointly trained to embed and retrieve, respectively, a bit sequence while minimizing the perceptual difference between the original audio and the watermarked signal.

The method differs from WavMark by training a detector instead of a decoder [21]. Additionally, a localization loss ensures that the detection of watermarked audio is performed at the level of individual samples. Thus, AudioSeal does not require synchronization bits or a computationally expensive brute-force search.

Special emphasis is placed on optimizing imperceptibility. To achieve this, the loss function includes Mel and Short-Time Fourier Transform multi-scale spectrogram terms to minimize the watermark's intensity. Additionally, a time-frequency loudness loss is employed, which leverages auditory masking psycho-acoustic properties [29], as the human auditory system fails perceiving weak and loud sounds occurring at the same time and at the same frequency range.

2.3. Classical signal processing attacks

In our study we consider some classical attacks often employed against watermarking techniques. Note that these are not exhaustive, as are only intended as a reference for our proposal, and comprise: a) addition of Gaussian noise at 25 and 15 dB, b) cutting and c) flipping random samples of the audio signal (at 0.1% and 0.5%), and d) low-pass and e) high-pass filtering with a butterworth filter of 16 coefficients (cut-off frequencies at 4kHz, 2kHz and 1kHz).

2.4. DNN-based SE attacks

We propose a novel attack for watermarking removal based on speech enhancement which leverages deep learning methods. An adversarial black-box approach is thus adopted, in which the DNN-based speech enhancement model is trained with stereo data, consisting of clean speech and their corresponding watermarked version. As a consequence, this approach reasonably assumes that the attacker has access to the watermarking service, but not to its internal workings. On the other hand, during inference, no assumptions are made: only the watermaked signal is required as input to the network, while its output is a watermark-removed signal.

As DNN SE architecture we consider the Deep Complex Convolutional Recurrent Network (DCCRN) from [30]. This network consists of a causal Convolutional Encoder-Decoder architecture with two Long Short-Term Memory (LSTM) layers between the encoder and the decoder, so that the temporal dependencies can be modeled. The DCCRN is essentially an extension of a Convolutional Recurrent Network (CRN) that performs a joint complex computation instead of considering two independent real and imaginary parts. Thus, the CNN and batch normalization layers in encoder/decoder are complex-valued, as well as the LSTM units. We chose this network because its ability to process full spectra (i.e., both magnitude and phase) obtained via Short-Time Fourier Transform (STFT) [31], defined as:

$$X(l,k) = \sum_{n=0}^{M-1} x(lH+n)w(n)e^{-j2\pi\frac{kn}{M}}, \quad (1)$$

where k is the frequency index (representing bins of $\frac{2\pi}{M}$ radians per sample), l is the frame index, $x(n)$ the is input signal, and $w(n)$ is the analysis window. As can be noted, $x(n)$ is segmented in overlapped frames of length M and shift H. In our proposal, we considered a square root Hann window for $w(n)$, and 512 and 256 samples for M and H, respectively.

Network's outputs in the form of a complex STFT, $\hat{X}(l,k)$, are finally reversed to the time-domain by means of the widely known Overlap and Add (OLA) method [32]. The resulting enhanced speech signal is then compared with the clean signal by the Scale-Invariant Signal-Distortion Ratio (SI-SDR), which evaluates the distortion in the time domain ignoring global gain or attenuation effects [33]. The ADAM optimization algorithm [34] with a batch size of 5 utterances, a learning rate of 10^{-3} and a dropout factor of 0.5 is used during training. Early stopping with a patience of 5 epochs (i.e., training is stopped after 5 epochs with no improvement in the validation set) is also applied to avoid over-fitting.

2.4.1. Double strike approach

Since speech enhancement deals with noise removal, we also considered a combined approach which mixes an additive noise attack with our proposal. In this scenario, the watermarked speech is first corrupted with noise and then enhanced using the DCCRN model. To train this model, which removes both noise and watermarks, we follow the previously described steps, but using watermarked signals with additive noise and clean speech as training pairs.

2.5. Performance metrics

We evaluate each attack on every watermarking (WM) technique. To this end, we consider the test dataset from each WM method (see Subsection 2.2). Performance metrics are then computed for each utterance and averaged across the dataset. These metrics aim to assess the robustness of the WM method, as well as the distortion, quality, and intelligibility of the resulting attacked signal.

Watermark robustness is measured in terms of the bit error rate (BER) between the extracted bit sequence (after the attack) and the original sequence associated with the utterance. This metric ranges from 0% to 100%, but a value of 50% or higher implies the complete destruction of the watermark. Distortion, quality, and intelligibility are assessed by comparing the attacked signal to the clean one (i.e., no watermark) using the SI-SDR [33], the Perceptual Evaluation of Speech Quality (PESQ) algorithm [25], and the Extended Short-Time Objective Intelligibility (ESTOI) [35], respectively. SI-SDR distortion is given in dB, PESQ uses a MOS scale from 1 (bad) to 5 (excellent), and ESTOI provides values within the interval $[0, 1]$. Unlike BER, for these metrics, higher values represent better performance. However, note that from the attacker's perspective, higher BER values indicate more successful attacks.

3. Results and discussion

Table 1 shows the performance results for the evaluated WM methods when they are attacked. For each specific WM method, the average BER, (SI-)SDR, PESQ, and ESTOI scores across the test dataset is provided per row, with each column representing a different attack. The last four rows provide a combined evaluation of all WM approaches. For classical attacks, this combined evaluation represents just the average performance metrics across all WM techniques. However, for the

WM technique	Performance metric	*No attack*	*Additive noise (25dB)*	*Additive noise (15dB)*	*Cut samples (0.1%)*	*Cut samples (0.5%)*	*Flip samples (0.1%)*	*Flip samples (0.5%)*	*Low pass (4kHz)*	*Low pass (2kHz)*	*High pass (1kHz)*	*High pass (2kHz)*	DCCRN	DCCRN + Noise (25dB)
DCT	BER (%)	0.0	0.8	15.1	0.1	0.0	0.1	0.5	7.2	13.9	**50.3**	50.1	28.3	34.5
	SDR (dB)	35.0	24.4	14.9	28.7	22.8	23.8	16.9	-16.4	-5.5	-5.9	-11.7	**39.9**	28.4
	PESQ	4.33	2.33	1.40	3.43	2.44	2.66	1.67	3.77	3.07	2.04	1.70	**4.52**	3.98
	ESTOI	0.994	0.975	0.877	0.989	0.974	0.979	0.938	0.991	0.778	0.343	0.183	**0.997**	0.986
Patchwork	BER (%)	0.0	3.2	6.9	1.7	6.2	5.1	12.8	3.8	7.0	0.0	12.8	15.5	**21.8**
	SDR (dB)	33.8	22.1	2.6	28.5	22.6	23.7	16.9	-2.1	-5.5	-5.9	0.3	**34.1**	27.5
	PESQ	4.14	2.04	1.06	3.35	2.41	2.62	1.66	3.58	3.02	1.89	1.66	**4.52**	3.94
	ESTOI	0.999	0.963	0.532	0.994	0.978	0.983	0.942	0.996	0.781	0.344	0.182	**0.999**	0.987
FSVC	BER (%)	0.0	0.3	0.7	0.0	0.0	0.1	0.6	3.3	3.0	8.5	**51.3**	15.0	17.8
	SDR (dB)	14.4	13.3	1.7	14.2	13.6	13.7	12.1	-15.3	-5.5	-6.1	-11.7	**14.9**	14.6
	PESQ	3.63	1.90	1.06	3.00	2.24	2.42	1.61	3.28	2.72	1.91	1.67	**3.96**	3.52
	ESTOI	0.966	0.931	0.516	0.961	0.946	0.950	0.912	0.963	0.755	0.335	0.182	**0.975**	0.964
Norm	BER (%)	7.2	7.2	7.3	7.2	7.2	7.2	7.2	12.8	13.9	15.7	17.3	**44.7**	43.9
	SDR (dB)	-4.2	**2.7**	-9.1	-4.3	-4.4	-4.4	-4.8	-3.3	-11.2	-18.8	-0.1	-0.8	1.2
	PESQ	**4.06**	2.01	1.06	2.89	2.22	2.34	1.59	3.04	2.53	1.76	1.56	3.27	2.74
	ESTOI	**0.989**	0.955	0.532	0.983	0.968	0.973	0.933	0.985	0.774	0.341	0.181	0.971	0.962
WavMark	BER (%)	0.0	2.4	22.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.6	46.0	**49.9**
	SDR (dB)	29.4	21.8	2.6	26.5	22.0	22.9	16.7	-16.4	-5.5	-6.0	-11.8	**30.9**	26.9
	PESQ	3.63	2.03	1.06	3.06	2.30	2.44	1.62	3.17	2.64	1.77	1.53	**4.10**	3.70
	ESTOI	0.983	0.954	0.535	0.979	0.965	0.969	0.932	0.980	0.774	0.340	0.180	**0.989**	0.982
AudioSeal	BER (%)	0.0	4.6	11.9	0.0	0.0	0.0	0.0	13.5	7.2	50.4	50.4	**78.5**	59.7
	SDR (dB)	27.9	25.0	8.5	25.8	21.7	22.6	16.6	-16.2	-5.4	-11.7	-11.7	**28.1**	25.3
	PESQ	4.42	2.77	1.16	3.45	2.46	2.67	1.67	3.78	3.04	1.76	1.76	**4.52**	3.90
	ESTOI	0.999	0.987	0.717	0.994	0.978	0.983	0.942	0.996	0.811	0.178	0.178	**0.999**	0.987
All	BER (%)	1.2	3.1	10.7	1.5	2.2	2.1	3.5	6.8	7.5	20.8	30.8	18.5	**36.6**
	SDR (dB)	22.7	18.2	3.5	19.9	16.4	17.0	12.4	-11.6	-6.4	-9.1	-7.8	**24.6**	19.6
	PESQ	4.04	2.18	1.13	3.20	2.35	2.52	1.64	3.44	2.84	1.85	1.65	**4.10**	3.72
	ESTOI	**0.988**	0.961	0.618	0.983	0.968	0.973	0.933	0.985	0.779	0.314	0.181	0.987	0.979

Table 1: *Performance results for each WM method (rows) when classical attacks and our proposed approach are applied (columns).*

DNN-based SE attacks (DCCRN and DCCRN+Noise), it involves a blind detection and removal of the watermark, independently of the method used, as training was conducted with clean-watermarked pairs from all the WM methods.

As can be observed, under no attack, the considered WM techniques achieve almost imperceptible marks and an outstanding intelligibility of the resulting speech. Only FSVC and WavMark methods yield PESQ scores under 4. Similarly, norm-space watermarking (Norm) provides very low SDR values, which suggest great changes in the waveform but with a scarce impact in perceptual quality and intelligibility. It also must be noted that this technique yields non-zero BERs, even without attacks, which can be explained by the silent segments in audio files (see 2.2.4).

Regarding the attacks, most of the classical ones are unable to remove the watermarks while, at the same time, introduce appreciable distortions within the signal. The most effective one is a high pass filtering with a cut-off frequency of 2 kHz, although some methods, such as WavMark and Patchwork, could deal with it. Nonetheless, high pass filtering is an extremely aggressive method which results in a perceived quality and intelligibility severely degraded. On the other hand, our proposed SE attack (DCCRN) is particularly effective against the DNN-based watermarking methods, WavMark and AudioSeal, (BER $\geq 46\%$) while significantly damages the payload of the classical ones (BER $\geq 15\%$). It also provides speech signals with an excellent quality and intelligibility. Indeed, for most of watermarking methods, perceived quality is even higher when the watermark is removed, which allows us for more aggressive attacks. Thus, at the cost of a small degradation of the quality, increased BERs can be achieved when the SE attack is combined with noise (DCCRN+Noise) as proposed in Section 2.4.1. As can be observed, our proposal provides the best trade-off between BER and signal degradation for all the considered WM techniques.

4. Conclusions

In this paper, we propose a novel watermarking attack based on speech enhancement. Our approach leverages a DCCRN architecture trained with pairs of clean and watermarked speech. The results demonstrate that our method can significantly degrade the embedded payload within the speech signal. For classical watermarking techniques, our attack yields BERs of at least 15%, while state-of-the-art DNN-based watermarking methods are completely obliterated. Moreover, we also measured the degradation, as well as the quality and intelligibility of the resulting watermark-removed speech, finding out that our method not only does not reduce these metrics but can even improve them. Thus, in this work, we demonstrate that DNN-based SE attacks pose a plausible and serious threat to current watermarking techniques.

5. Acknowledgements

This paper is part of the project PID2022-138711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

6. References

[1] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, 2021, pp. 5530–5540.

[2] E. Casanova *et al.*, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 17–23 Jul 2022, pp. 2709–2720.

[3] C. Wang *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[4] J. A. Gonzalez *et al.*, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[5] J. A. Gonzalez-Lopez *et al.*, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[6] Forbes, "Fraudsters cloned company director's voice in $35 million bank heist, police find," 2024. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions

[7] Avast, "Voice fraud scams company out of 243,000 dollars," 2024. [Online]. Available: https://blog.avast.com/deepfake-voice-fraud-causes-243k-scam

[8] Z. Wu *et al.*, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[9] X. Liu *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[10] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1985–1999, 2019.

[11] E. Rosello, A. Gomez-Alanis, A. M. Gomez, and A. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Proc. INTERSPEECH 2023*, 2023, pp. 5281–5285.

[12] S. H. Mun *et al.*, "Towards single integrated spoofing-aware speaker verification embeddings," in *Proc. INTERSPEECH 2023*, 2023, pp. 3989–3993.

[13] E. Parliament, "European artifical intelligence act," 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

[14] H.-T. Hu and L.-Y. Hsu, "Robust, transparent and high-capacity audio watermarking in dct domain," *Signal Processing*, vol. 109, pp. 226–235, 2015.

[15] I. Natgunanathan *et al.*, "Patchwork-based multilayer audio watermarking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2176–2187, 2017.

[16] J. Zhao *et al.*, "Desynchronization attacks resilient watermarking method based on frequency singular value coefficient modification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2282–2295, 2021.

[17] S. Saadi, A. Merrad, and A. Benziane, "Novel secured scheme for blind audio/speech norm-space watermarking by arnold algorithm," *Signal Processing*, vol. 154, pp. 74–86, 2019.

[18] G. Hua *et al.*, "Twenty years of digital audio watermarking—a comprehensive review," *Signal Processing*, vol. 128, pp. 222–242, 2016.

[19] C. Liu *et al.*, "Dear: A deep-learning-based audio re-recording resilient watermarking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 201–13 209.

[20] G. Chen *et al.*, "WavMark: Watermarking for audio generation," 2024. [Online]. Available: https://arxiv.org/abs/2308.12770

[21] R. S. Roman *et al.*, "Proactive detection of voice cloning with localized watermarking," 2024. [Online]. Available: https://arxiv.org/abs/2401.17264

[22] C. Qin *et al.*, "Visible watermark removal scheme based on reversible data hiding and image inpainting," *Signal Processing: Image Communication*, vol. 60, pp. 160–172, 2018.

[23] M. W. Hatoum, J.-F. Couchot, R. Couturier, and R. Darazi, "Using deep learning for image watermarking attack," *Signal Processing: Image Communication*, vol. 90, p. 116019, 2021.

[24] J. Garofolo *et al.*, "The structure and format of the DARPA TIMIT CD-ROM Prototype," *in Proceedings of NIST*, 1988.

[25] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recomendation P.862*, 2000.

[26] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, no. 4, p. 1423 – 1443, 2001.

[27] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, S. Bengio *et al.*, Eds., vol. 31. Curran Associates, Inc., 2018.

[28] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," 2022. [Online]. Available: https://arxiv.org/abs/2210.13438

[29] J. Schnupp, I. Nelken, and A. J. King, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, 11 2010.

[30] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv:2008.00264*, 2020.

[31] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[32] R. E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 1, p. 99–102, 1980.

[33] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of 3rd International Conference on Learning Representations*, 2015, pp. 1–13.

[35] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.