MASKMARK: ROBUST NEURAL WATERMARKING FOR REAL AND SYNTHETIC SPEECH

Patrick O'Reilly[1], *Zeyu Jin*[2], *Jiaqi Su*[2], *Bryan Pardo*[1]

[1]Northwestern University [2]Adobe Research

ABSTRACT

High-quality speech synthesis models may be used to spread misinformation or impersonate voices. Audio watermarking can combat misuse by embedding a traceable signature in generated audio. However, existing audio watermarks typically demonstrate robustness to only a small set of transformations of the watermarked audio. To address this, we propose MaskMark, a neural network-based digital audio watermarking technique optimized for speech. MaskMark embeds a secret key vector in audio via a multiplicative spectrogram mask, allowing the detection of watermarked speech segments even under substantial signal-processing or neural network-based transformations. Comparisons to a state-of-the-art baseline on natural and synthetic speech corpora and a human subjects evaluation demonstrate MaskMark's superior robustness in detecting watermarked speech while maintaining high perceptual transparency.

Index Terms— Watermarking, speech synthesis, synthetic media

1. INTRODUCTION

Generative speech models have numerous beneficial applications, such as aiding individuals with speech difficulties and fostering creative expression. However, they can also be used for malicious purposes, including voice impersonation and disseminating false information. Recent speech synthesis systems allow users to "deepfake" the voices of public figures to make statements ranging from incongruous and comedic to offensive and hateful [1, 2]. Voice conversion has been used to replicate the voices of famous singers [3], igniting debates on copyright and intellectual property. Moreover, the prevalence of realistic synthetic media has been used in attempts to cast doubt on the authenticity of digital evidence in court [4].

Developers and providers of generative speech models therefore need methods to identify speech produced by their models, while preserving the sonic quality of the generated speech, thereby ensuring that creative applications are not hindered. Existing classification-based approaches to synthetic speech detection leverage neural network systems trained on corpora of synthetic speech to recognize characteristic artifacts [5, 6]; however, such approaches often fail to generalize to unseen synthesis systems. Moreover, the distribution of synthetic speech is constantly evolving: many once-characteristic synthesis artifacts (e.g. aliasing and pitch inconsistency) have been mitigated through architectural advances [7, 8], and recent systems achieve results comparable to natural speech in human listening tests [9, 10]. As speech synthesis technology improves, synthetic speech classification systems will likely face an increasingly difficult task.

One promising alternative is digital audio watermarking, where an embedding algorithm conceals a traceable signature (*key*) in audio signals and a corresponding detection algorithm determines whether the key is present in an audio signal. This allows the watermarked output of a speech synthesis system to be distinguished from natural speech, even in the absence of other identifying artifacts [11].

To enable practical synthetic speech detection, a watermark should satisfy two criteria. First, to avoid compromising the quality of synthetic speech, the watermark should be *perceptually transparent* – imperceptible to human listeners. Second, the watermark should be *robust* – capable of operating with strong detection accuracy even when the watermarked audio is significantly altered. Such alterations may include passive transformations of the watermarked audio from transmission over digital and physical channels (e.g. equalization, reverberation, codec compression); active transformations that occur through benign user interactions (e.g. processing synthetic speech clips to remove pauses); and transformations specifically intended to thwart watermark detection (e.g. pitch/time-scale modification).

Existing speech watermarks struggle to provide robust discrimination between watermarked and unwatermarked speech while maintaining the requisite perceptual transparency for widespread use. Tai & Mansour propose an autocorrelation-based spread-spectrum watermark for suppressing wake-word detection on Amazon Alexa [12]. While their proposed "Eigen" watermark achieves state-of-the-art performance in discriminating between short watermarked and unwatermarked speech segments in realistic acoustic environments, the watermark is vulnerable to certain transformations such as pitch- and time-scale modification.

Others have sought to leverage the strengths of *deep neural networks* for watermarking audio in a data-driven manner. Liu et al. propose training paired embedder and detector networks to perform music watermarking in the wavelet domain [13]. However, their method requires long audio segments for embedding and produces clearly audible artifacts when applied to speech. Pavlović et al. also propose training paired networks to watermark short speech segments [14]. While their "DNN-A" system demonstrates strong perceptual transparency, it is not robust to common editing distortions. Moreover, we find that the low-capacity training scheme employed by Pavlović et al. results in the detector "overfitting" to a small set of key vectors, producing unacceptably high false-positive rates when discriminating between watermarked and unwatermarked speech.

To address these shortcomings we propose MaskMark, a robust neural network-based watermarking method for speech audio. We compare MaskMark to Eigen and DNN-A, the strongest signal-processing and neural network-based speech watermarks, respectively, of which we are aware. Our contributions are as follows:

- We design a novel network architecture for embedding watermarks in short ($\leq$ 2s) speech segments via a multiplicative spectrogram mask, and build on the method of Pavlović et al. to significantly improve robustness and discriminative performance.
- We show that MaskMark achieves superior robustness to Eigen and DNN-A under a comprehensive set of audio transformations encompassing editing manipulations, channel

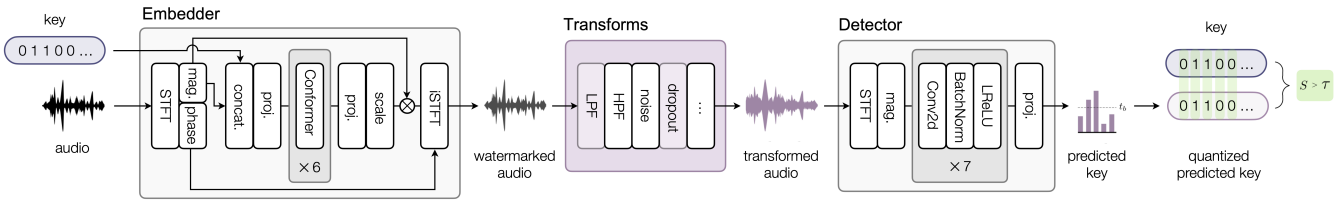

Fig. 1. The proposed MaskMark architecture.

effects, and attacks designed to remove watermarks.

- Through objective and subjective evaluations, we show that MaskMark maintains perceptual transparency competitive with or better than Eigen, the only baseline method to demonstrate practical performance. [1]

2. WATERMARKING METHOD

We propose a neural network-based method for watermarking speech audio. Building upon the approach of Pavlović et al. [14], we use a two-network system to embed and detect watermarks, where each watermark is represented as a binary key vector of length 512, drawn from a fixed pool of n keys. An *embedder* network conceals a watermark key within a given *carrier* audio recording, while a *detector* network recovers the watermark by predicting the key vector contained in the carrier. The networks are trained such that the vector predicted by the detector closely matches the key embedded within a watermarked carrier but is essentially random for unwatermarked audio. Thus, given knowledge of the embedded key, the operator of the detector can distinguish between watermarked and un-watermarked audio by comparing any predicted vectors to the known key via a calibrated similarity threshold.

2.1. Architecture

Embedder: Our novel embedder network takes as input a key vector and carrier signal to be watermarked. To allow for embedding in high sample-rate audio, the carrier is resampled to 48kHz and a magnitude spectrogram is computed with a frame length of 1024 samples, hop length of 512 samples, and Hann window. The key vector is repeated and concatenated with each frame, and the result projected to a hidden dimension of 576 channels. This embedded representation then passes through six Conformer [15] layers with feed-forward dimension 768, six attention heads, and kernel size 5. The output is projected to match the frequency dimension of the carrier spectrogram and bounded to $\left[1\mathrm{e}^{-5}, \infty\right)$ to obtain a mask. After multiplying the mask and carrier magnitude spectrogram, we take the inverse Fourier transform using the original carrier phase to obtain the watermarked waveform. In total, the embedder network contains 14.3m parameters. Whereas DNN-A directly generates the complex spectrogram of the watermarked audio, we find that the use of a real-valued spectrogram mask to embed the watermark allows our approach to easily generalize across input sample rates and prevents the embedder from introducing noise in silent regions of the carrier, as the masking operation is multiplicative. Additionally, the redundant frame-level embedding and Conformer module allow our embedder to robustly encode a larger set of key vectors and thereby avoid the "overfitting" behavior described in Section 3.2.

[1] We provide audio examples at https://interactiveaudiolab.github.io/project/maskmark.html

Detector: We adopt the fully-convolutional detector of Pavlović et al. with minor modifications. We prepend an additional strided convolutional layer, resulting in a receptive field of approximately one second at 48kHz (versus the original two seconds at 16kHz), and operate the detector on the same magnitude spectrogram representation as the embedder rather than a complex spectrogram. The detector predicts a 512-dimensional vector of unnormalized probability scores v, where v_i represents the probability that the i^{th} bit in the watermark key is 1. At inference time, we quantize detector outputs to $\{0, 1\}^{512}$ using a fixed threshold of $t_b = 0.5$. In total, our detector network contains 5.1m parameters.

Scoring: As in previous works, we assume the detector operator has access to a watermark key [12, 14, 16]. Given candidate audio, we compute the proportion of matching bits between the detector's quantized output and the known key to obtain a score S that is then compared to a calibrated threshold τ. If $S \geq \tau$, we conclude the candidate audio contains the watermark. We find this outperforms other scoring methods such as cosine distance, mean squared error, and binary cross-entropy.

2.2. Training

While Pavlović et al. use a private speech dataset, we train on the VCTK dataset [17], totalling 44 hours of audio across 108 speakers sampled at 48kHz. We train on 2-second random excerpts for $70,000$ iterations with batch size 64 and mixed precision. We use a fixed set of n=64 random watermark keys rather than the n=6 used by DNN-A, as we find this prevents the detector from "overfitting" to the key set and encourages more uniform random predictions over unwatermarked audio (see Section 3.2).

At each training iteration, we energy-normalize carrier audio to -24dBFS, apply a random phase shift, and pass the carrier to the embedder network alongside a randomly selected key vector to obtain watermarked audio. Watermarked audio is then passed through a differentiable channel simulation. While past works have focused on additive noise, pass filters, and reverberation [13, 14], we also consider pitch- and time-scale modification, as such transformations often prove especially destructive to watermarks. Our simulation encompasses low- and high-pass filtering at 4000Hz and 500Hz, respectively; recorded environmental noise [18, 19, 20] at 20-35 dBSNR; sample dropout with probability 0.001; reverberation applied via recorded and simulated impulse responses [19, 21]; phase-vocoder pitch shift of up to ± 1 semitone; speed change of $\pm 5\%$; and spectral gating applied via a threshold relative to the highest-magnitude bin of each spectrogram frame [22]. We find that training on this relatively small set of transformations allows MaskMark to generalize to a much broader set of unseen conditions (see Section 3.2).

At each training step, we randomly select three of the aforementioned transformations and apply each with an independent probability of 25%. Transformed audio is then passed to the detector network, which predicts an unnormalized probability distribution over

the key vector. We compute two losses: binary cross-entropy loss on the detector predictions and embedded watermark keys; and a multi-resolution spectrogram loss on the embedder output and input [23] to encourage perceptual transparency. Gradients of the cross-entropy loss are propagated to both the detector and embedder (through the differentiable channel simulation), while gradients of the spectrogram loss are propagated only to the embedder. We balance the respective norms of these gradients within the embedder in a ratio of $\frac{100}{1}$ using the method of Défossez et al. [24] before performing an update. To ensure stability, we use AutoClip [25] to perform adaptive gradient clipping in both the embedder and detector networks at the 10^{th} percentile. We use the NAdam optimizer [26] with learning rate $2\text{e}{-4}$, β_1=0.9, β_2=0.999, and weight decay $1\text{e}{-6}$.

3. EXPERIMENTS

We perform experiments (1) validating MaskMark's ability to robustly discriminate between watermarked and unwatermarked audio and (2) measuring its perceptual transparency and objective quality metrics. We evaluate on two datasets unseen in training. To represent current synthesis systems, we use the VITS text-to-speech model [9] to resynthesize utterances from 10 unseen VCTK speakers at 24kHz; and to represent high-quality future synthesis systems, for which generated speech may be "indistinguishable" from natural speech, we use the natural speech DAPS "clean" subset [27], containing studio recordings of 20 speakers sampled at 44.1kHz.

3.1. Baselines and comparison methodology

Eigen: We implement the Eigen watermark of Tai & Mansour [12] using an embedding band of 3-4kHz, non-overlapping windows of length 100ms, segment length $N_s = 2$, and repetition $N_r = 50$ to obtain a watermark length of one second as in the original work. Because the watermark strength parameter β is not specified in the original work, we evaluate two configurations with $\beta = 1.0$ and $\beta = 0.5$ to balance perceptual transparency and robustness.

DNN-A: We implement the DNN-A watermark of Pavlović et al. [14] using the hyperparameters detailed in the original work, including the differentiable transform set, and train on 2-second excerpts of the VCTK dataset for $140,000$ iterations to ensure convergence. Our implementation achieves comparable robustness (as measured by key recovery under the dropout, filtering, and noise transforms considered in the original work) and perceptual transparency (as measured by PESQ [28] and SNR) to the authors' implementation.

Conflicting sample rates: MaskMark operates at 48kHz, Eigen at 44.1kHz, and DNN-A at 16kHz. For MaskMark and Eigen, we account for sample rate differences by resampling audio before and after embedding. For DNN-A, we split carrier audio into two bands at 8kHz before embedding. The low band is resampled to 16kHz, embedded, and resampled to the original dataset rate (44.1kHz for DAPS, 24kHz for VITS). We then balance the energy of the two bands to match their pre-embedding ratio and sum. We find this does not impact robustness, and allows for fair listening comparisons between watermarking methods at the original dataset rates.

Watermark length: Following Pavlović et al., we perform robustness evaluations at a watermarked segment length of 2 seconds. The proposed MaskMark is capable of embedding at arbitrary lengths due to redundant frame-wise processing, and we simply repeat the Eigen watermark. For MaskMark and Eigen, which render detection scores at 1-second intervals, we take the maximum over all scores produced for a segment. Finally, for our perceptual evaluations we embed repeatedly to cover entire utterances.

3.2. Robustness to audio transformations

We evaluate the robustness of the proposed and baseline watermarks in a synthetic speech detection scenario. For each of the aforementioned watermarks and datasets, we sample 2000 two-second excerpts and embed with random keys to generate pairs of watermarked and unwatermarked utterances. We then apply an audio transformation (see below) and compute detector scores for the transformed utterances using the known keys. For Eigen, we compute the self-correlation score as defined by Tai & Mansour [12]. For both MaskMark and DNN-A, we compute the bitwise similarity score described in Section 2.1 to allow for fair comparisons. For a watermark to be of practical use, the distributions of synthetic (watermarked) and natural (unwatermarked) scores should be sufficiently separated to allow for reliable detection with a low false-positive rate; following previous work in synthetic media detection [29], we compute the achievable true positive rate when the threshold τ is empirically calibrated to fix the false positive rate at 1.0% (TPR@1%FPR).

We evaluate the robustness of the proposed and baseline methods over a set of 26 realistic audio transformations, including 18 unseen in training. **Signal-processing**: we modify the training transformations listed in Section 2.2 to use unseen impulse responses, noise recordings, and speed change ratios within the interval ±5%; we also lower noise SNR to 10dB. For evaluation we implement the following additional transformations: Mu-Law quantization with 256 channels; dynamic range compression [30]; constant phase shift in $\pm\pi$; clipping applied at the 95^{th} percentile of observed amplitudes; OGG-Vorbis compression; equalization across six bands, each with randomly sampled gain in ±1dB; silence removal; and phase-vocoder time-stretching in the interval ±10%. **Neural vocoding**: we transform audio by resynthesizing from mel-spectrograms using the pre-trained neural vocoder models BigVGAN [7], Vocos [31], DiffWave [32], and HiFiGAN [33]. The mel-spectrogram representation discards a large amount of information from the audio, which is subsequently inferred by the vocoder. As a result, this transformation may substantially alter or destroy embedded watermarks while maintaining much of the audio's perceptual quality. **Neural codec compression**: we transform watermarked audio by applying the state-of-the-art neural codec models Descript Audio Codec (DAC) [34] and Encodec [24]. Similar to vocoding, neural network-based compression reconstructs audio from a tight information bottleneck and may degrade the effectiveness of embedded watermarks. **Neural speech enhancement** (DNN SE): speech enhancement models remove or suppress non-speech sounds from a recording; thus, if an embedded watermark can be distinguished from the carrier speech (e.g. if it manifests audible artifacts), it may be partially or wholly removed. We use pretrained Demucs [23] and MetricGAN+ [35] models.

Experimental results are in Figure 2. We find that while DNN-A often recovers embedded key vectors with high accuracy, the method can not distinguish between watermarked and unwatermarked audio at sufficiently low false-positive rates to be of practical use. We hypothesize training on n=6 keys causes the detector network to "overfit"; as a result, on average 4% of *unwatermarked* utterances are mapped extactly to key vectors (and thus perfect detector scores) even in the absence of any transformations – an impractical lower bound on the operating false positive rate.

By contrast, MaskMark achieves competitive or better robustness under every evaluated transformation when compared to Eigen.

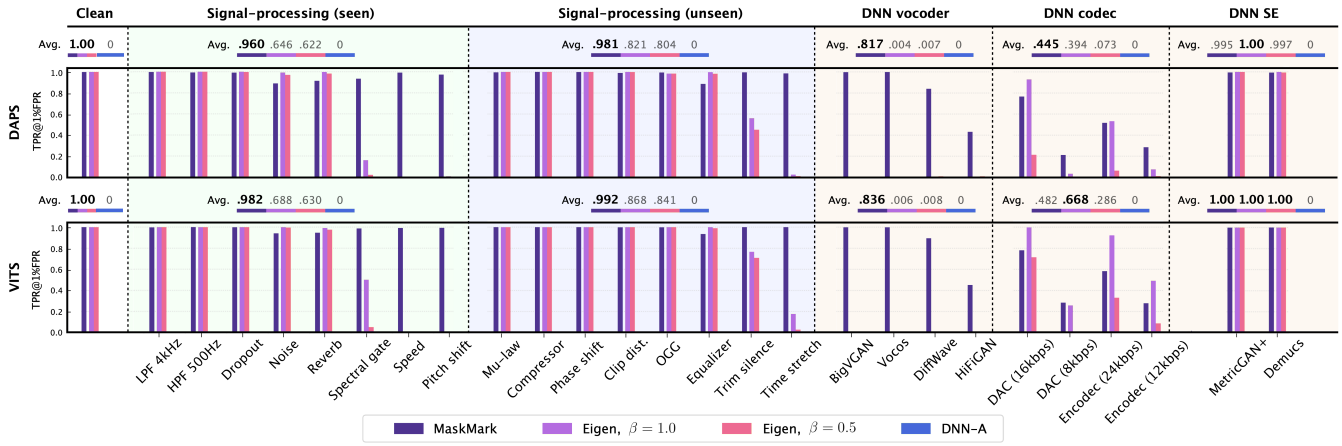

Fig. 2. TPR@1%FPR for watermarks under selected transformations, indicating the robustness of each method when configured to operate with a fixed 1% false-positive rate. Scores in the top row are for the DAPS dataset, scores in the bottom row for our VITS dataset. MaskMark (proposed) is robust to more transformations than the baselines. DNN-A has a true positive rate of 0% for false-positive rates below 4%.

MaskMark performs significantly better in the average and worst case over each set of transformations with the exception of neural speech enhancenent (in which both Eigen and MaskMark show near-perfect performance) and neural codec compression, for which Eigen performs better on the VITS dataset in the average case. MaskMark demonstrates strong robustness to transformations for which other methods fail completely at low false positive rates (time-stretching, neural vocoding), and generalizes far outside its relatively small training dataset and limited training distribution of signal-processing transformations.

3.3. Perceptual transparency

We assess the perceptual transparency of MaskMark through standard speech quality metrics and a human listener study. Following prior work [14], we compute SNR and PESQ [28] scores over 500 pairs of watermarked and unwatermarked utterances on each dataset. We sample 5-second excerpts from the DAPS dataset and full recordings from the VITS dataset to ensure sufficiently long utterances. Additionally, we conduct a human listening study on Prolific [2] in which 119 participants are asked to distinguish between pairs of watermarked and unwatermarked utterances. In this format, a 50% identification rate corresponds to a perfectly transparent watermark, while higher identification rates correspond to diminished transparency. We select 50 utterance pairs from each dataset, yielding a total of n=800 trials per method. Each utterance is assigned to 16 different participants from the United States. The participants are required to wear headphones and pass a listening test to ensure they can distinguish subtle differences in audio samples. Each assignment covers 4 utterances processed by all testing methods and 4 secret validation tests to ensure participants are paying attention. Participants can complete as many unique assignments as they please, and are paid at $15 per hour. Table 1 shows the results of our evaluations. .

While MaskMark produces the lowest average SNR, we note that SNR is designed to measure the magnitude of additive perturbations of the original signal, while our method produces highly correlated filtering-like perturbations through the use of spectral masking. This lets MaskMark introduce large-magnitude perturbations while also maintaining the highest PESQ scores of all evaluated methods.

In the human evaluation, we find that all methods, save DNN-A, are distinguishable from clean audio at $p < 0.001$. While DNN-A obtains the best perceptual transparency, this comes at the expense of detection performance too poor for practical use: it fails to detect true watermarks when the false positive rate is $< 4\%$. We note that, even when given a clean reference and explicitly instructed to listen for the watermark, listeners are only able to correctly distinguish between MaskMark and clean audio roughly 60% of the time, where chance performance is 50%. Via the Chi-Squared independence test, Eigen (β=1.0) is significantly more perceptible than the proposed method at $p << 0.001$, while Eigen (β=0.5) does not differ significantly ($p = 0.3057$ for DAPS and $p = 0.918$ for VITS). Overall, MaskMark is not significantly more perceptible than the "weak" Eigen configuration (β=0.5) while achieving superior robustness to the "strong" Eigen configuration (β=1.0).

	SNR↑		PESQ↑		Perceptual ID↓	
Speech dataset	DAPS	VITS	DAPS	VITS	DAPS	VITS
DNN-A [14]	30.06	38.43	4.39	4.38	51.4%	51.2%
Eigen [12], $\beta = 1.0$	23.40	21.10	4.24	4.12	78.2%	76.3%
Eigen [12], $\beta = 0.5$	28.82	26.43	4.44	4.40	61.4%	60.5%
MaskMark	19.55	22.01	4.56	4.59	63.9%	60.8%

Table 1. Objective speech quality metrics and human identification rates. DNN-A values are shaded to indicate it fails to detect watermarked speech at low false-positive rates ($< 4\%$).

4. CONCLUSION

In this work we propose MaskMark, a novel neural network-based watermark for synthetic speech detection. Experimental results show that MaskMark can discriminate between watermarked and unwatermarked speech more robustly than state-of-the-art signal-processing and neural network-based speech watermark baselines, while matching or exceeding the perceptual transparency of the only baseline to achieve nontrivial robustness. Although neural network-based transformations (vocoder resynthesis and neural codec compression) are challenging for watermarks, the proposed system proves more robust to these transformations than the baselines. Future work may incorporate such transformations directly into training to improve robustness, or explore the effectiveness of stronger optimization-based adversarial attacks on watermark detectors.

[2] https://www.prolific.co/

5. REFERENCES

[1] S. Dowell, "Polish startup wows with 'deep fake' voice cloning tool which can make leonardo dicaprio sound like kim kardashian," *TheFirstNews*, 2023.

[2] M. Moon, "A new ai voice tool is already being abused to make deepfake celebrity audio clips," *Engadget*, 2023.

[3] J. Coscarelli, "An a.i. hit of fake 'drake' and 'the weeknd' rattles the music world," *The New York Times*, 2023.

[4] S. Bond, "People are trying to claim real videos are deepfakes. the courts are not amused," *National Public Radio*, 2023.

[5] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "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.

[6] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023.

[7] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," in *ICLR*, 2023.

[8] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive gan for conditional waveform synthesis," in *ICLR*, 2022.

[9] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, 2021.

[10] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, F. Soong, T. Qin, S. Zhao, and T. Liu, "Naturalspeech: End-to-end text to speech synthesis with human-level quality," *arXiv preprint arXiv:2205.04421*, 2022.

[11] Y. Cui, L. He, and S. Zhao, "Introducing the watermark algorithm for synthetic voice identification," *Azure AI services Blog*, 2022.

[12] Y. Tai and M. F. Mansour, "Audio watermarking over the air with modulated self-correlation," in *ICASSP*, 2019.

[13] C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang, and N. Yu, "Dear: A deep-learning-based audio re-recording resilient watermarking," in *AAAI*, 2023.

[14] K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski, "Robust speech watermarking by a jointly trained embedder and detector using a dnn," *Digital Signal Processing*, vol. 122, pp. 103381, 2022.

[15] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[16] D. Kirovski and H.S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1020–1033, 2003.

[17] J. Yamagishi, "English multi-speaker corpus for cstr voice cloning toolkit," 2012.

[18] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[19] T. Ko, V. Peddinti, D. Povey, M. L Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017, pp. 5220–5224.

[20] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Interspeech*, 2019.

[21] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[22] P. O'Reilly, A. Bugler, K. Bhandari, M. Morrison, and B. Pardo, "Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models," in *Neural Information Processing Systems*, 2022.

[23] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[24] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[25] P. Seetharaman, G. Wichern, B. Pardo, and J. Le Roux, "Autoclip: Adaptive gradient clipping for source separation networks," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020.

[26] T. Dozat, "Incorporating nesterov momentum into adam," *ICLR Workshop*, p. 2013–2016, 2016.

[27] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? - a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, 2015.

[28] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[29] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust," *arXiv preprint arXiv:2305.20030*, 2023.

[30] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style transfer of audio effects with differentiable signal processing," *Journal of the Audio Engineering Society*, 2022.

[31] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," *arXiv preprint arXiv:2306.00814*, 2023.

[32] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *ICLR*, 2021.

[33] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[34] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *arXiv preprint arXiv:2306.06546*, 2023.

[35] S. Fu, C. Yu, T. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," in *Interspeech*, 2021.