

CLONEBLOCK-RT: STREAMING VOICE PROTECTION AGAINST ZERO-SHOT VOICE CLONING IN LIVE TELEPHONY

Yuxuan Liu, Kunyang Zhang, Peihong Zhang, Yiqiang Cai, Yizhou Tan, and Shengchen Li

Xi'an Jiaotong-Liverpool University, Suzhou, China

ABSTRACT

Zero-shot voice cloning can synthesize speech in an unseen speaker's voice from only a short reference recording, creating a risk that speech captured during a call may be reused for speaker impersonation. Proactive protection can reduce the usefulness of exposed speech as a cloning reference, but live calls require the defense to operate causally as speech is produced. In this setting, real-time execution alone is insufficient: the protected signal must remain effective after telephone processing, while an attacker can later select and combine multiple recorded segments as cloning references. We present CloneBlock-RT, a single-pass causal STFT gain predictor trained to reduce speaker-embedding similarity after matched simulated telephone transformations while constraining spectral modification. We evaluate CloneBlock-RT under session-level reference collection, synthesis-based voice cloning, speech-utility, and streaming-runtime tests. The protection retains measurable reductions in speaker similarity when multiple same-call segments are selected and combined, and these reductions persist in the resulting cloned speech rather than being limited to an embedding-level proxy. The defense introduces modest speech-utility trade-offs while remaining compatible with real-time processing. Together, these results indicate that causal, telephone-aware protection can remain effective after transmission and retrospective multi-segment reference construction under the evaluated conditions. Code is available at: https://anonymous.4open.science/r/CloneBlock_RT-FB43.

Index Terms— Zero-shot voice cloning, Proactive voice protection, Speech privacy, Streaming signal processing, Telephony

1. INTRODUCTION

Zero-shot voice cloning can synthesize speech in the voice of an unseen speaker from only a short reference recording [1, 2]. This capability creates a direct audio-privacy risk: speech originally shared for communication can later be reused as a cloning reference for speaker impersonation. Proactive voice protection addresses this threat by modifying speech before it becomes available to a cloning system, with the goal of reducing speaker-identity information while preserving the linguistic content and usability of the original signal [3, 4, 5]. For prerecorded audio, such protection can be applied after the whole utterance is available. Live telephone calls impose a substantially different constraint: protection must be applied synchronously as speech is produced, before the defender knows which parts of the conversation an attacker may later reuse.

Existing work has begun to address this real-time requirement. Some proactive defenses are designed primarily for prerecorded speech, where complete utterances are available during protection [6, 7]. Streaming variants and real-time defenses, including VS-Mask [3], LiveMask [4] and Enkidu [8], show that low-cost online

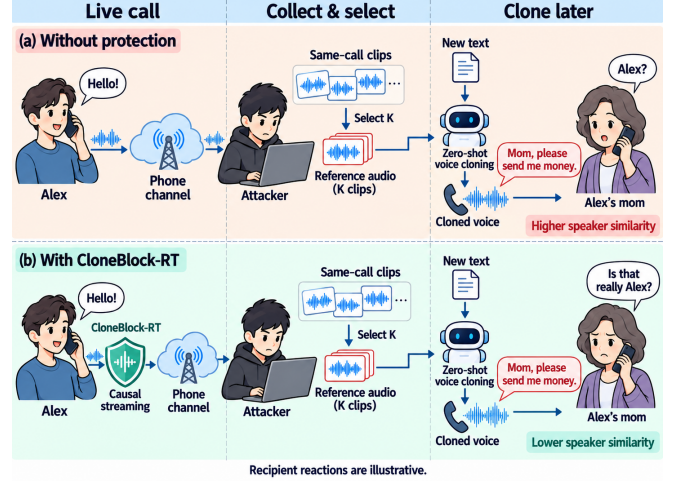


Fig. 1. Live-call threat model: (a) unprotected speech is collected for later zero-shot voice cloning; (b) CloneBlock-RT causally protects outgoing speech before telephone transmission to reduce speaker similarity.

voice protection is feasible. These approaches address an important deployment bottleneck: the protection transform can be applied without repeatedly solving a costly optimization problem for every incoming utterance. For live telephony, however, computational feasibility is only one part of the problem. The protected waveform is subsequently altered by the communication channel, and the attacker can retain the resulting recording and decide which speech segments to use only after the call has finished. A useful live-call defense must therefore remain effective not only during causal processing, but also after telephone transmission and subsequent reference construction.

The first technical challenge is therefore to make protection *channel-aware* while retaining causal, low-cost operation. Telephone transmission can alter speech through band-limiting, resampling, automatic gain control, and noise, and may additionally involve codec processing such as Opus [9] and G.711 [10]. A perturbation that changes speaker embeddings before transmission need not produce the same identity change in the signal eventually recorded by the attacker. At the same time, increasing the perturbation magnitude is not a satisfactory solution because speech intelligibility and conversational quality must be preserved. The resulting design problem is to suppress speaker-discriminative information that survives the communication channel while constraining the online modification to a lightweight and content-preserving signal transformation.

A second challenge arises after the call has been recorded. Unlike an utterance-level evaluation in which a defense is tested on one

predefined segment, a telephone recording provides many possible reference clips. An attacker can select several segments and use them jointly as a reference, as illustrated in Fig. 1. Consequently, a similarity reduction measured on an isolated window does not by itself establish how much speaker information remains in a multi-segment reference. More generally, real-time protection and reference construction occur at different stages: the defender must modify each frame before knowing which segments will later be selected, whereas the attacker can make this decision retrospectively. This motivates evaluating streaming protection at the session level rather than only on independent utterances.

We address these coupled requirements with **CloneBlock-RT**, which optimizes speaker-identity suppression *after simulated telephone processing* while restricting inference to a single-pass causal transformation. During training, protected and unprotected branches undergo matched telephone transformations, with their post-channel speaker embeddings defining the identity-reduction objective. At inference, a lightweight causal predictor produces bounded frequency-band gains for each arriving short-time Fourier transform (STFT) frame. Multiplicative gains preserve phase, while reconstruction and energy penalties constrain speech modification during training. The resulting protection is input-dependent and requires one forward pass without iterative test-time optimization.

Our evaluation addresses both telephone transmission and subsequent reference construction. We construct pseudo-call sessions from CSTR VCTK, an English multi-speaker corpus covering 110 speakers with diverse accents [11], and apply matched telephony processing to protected and unprotected streams. Using ECAPA-TDNN embeddings [12], we measure speaker similarity when same-session windows are selected and pooled across reference sizes K . Targeted greedy selection uses the unprotected, telephone-processed session embedding as an oracle target, forming an oracle-assisted stress test. We then assess whether these embedding-level reductions persist in speech generated by XTTS-v2, a multilingual zero-shot text-to-speech system supporting one or multiple reference recordings [13]. Speech-recognition, intelligibility, and runtime measurements quantify the utility and computational cost of streaming protection.

The main contributions of this work are:

- **Telephony-aware streaming protection:** We introduce CloneBlock-RT, a causal single-pass spectral gain predictor whose speaker-identity objective is evaluated after matched simulated telephone processing, enabling input-dependent protection without iterative optimization during a call.
- **Session-level multi-reference evaluation:** We evaluate protection beyond isolated windows by combining selected-window embedding pooling with multi-clip XTTS cloning experiments, and jointly report speaker-similarity reduction, speech utility, and streaming runtime under the evaluated telephony conditions.

2. CLONEBLOCK-RT: METHOD AND EVALUATION

Figure 2 shows protection before telephone transmission and subsequent reference collection. LiveMask and Enkidu reuse precomputed filters or perturbations for real-time processing [4, 8]. CloneBlock-RT instead predicts frame-dependent gains trained against a matched post-channel identity objective.

2.1. Telephony-aligned training

A similarity reduction before transmission need not survive telephone processing. For example, a linear band-limiting operator H maps a

modification $\delta = \tilde{x} - x$ to $H\delta$; components removed by H cannot affect the received speech. We therefore train the protector using similarity measured after simulated transmission.

Let $\tilde{x} = \mathcal{D}_\phi(x)$, where ϕ denotes the trainable protector parameters. A channel configuration θ , sampled from the training distribution P_{tr} , specifies telephone processing such as band-limiting, resampling, gain control, and noise. For each sampled configuration, we process both the original and protected speech under the same channel settings. With a fixed speaker encoder h and L2 normalization $\nu(v) = v/\|v\|_2$, define

$$\begin{aligned} y_\theta^{(c)} &= \mathcal{T}(x; \theta), & y_\theta &= \mathcal{T}(\tilde{x}; \theta), \\ \ell_\theta(x, \phi) &= \nu(h(y_\theta^{(c)}))^\top \nu(h(y_\theta)), \\ \mathcal{L}_{\text{spk}}(x, \phi) &= \mathbb{E}_{\theta \sim P_{\text{tr}}} [\ell_\theta(x, \phi)]. \end{aligned} \quad (1)$$

Here (c) denotes the unprotected branch. The loss ℓ_θ is the cosine similarity under one channel configuration; matching the settings controls for differences in channel processing between the two branches.

The expectation averages these scalar losses over the training channel distribution, not the waveforms or embeddings. It is estimated through channel sampling during training: each sampled loss provides an update signal for the protector, while h remains fixed. This follows the expectation-over-transformation principle of optimizing an objective over transformed inputs [14]. Sampling different conditions encourages similarity reduction that persists across the modeled channels, rather than relying on a modification effective only before transmission. This is an average-case objective, not a guarantee for every channel. Channel simulation and loss computation are used only during training and add no operations to the deployed protector.

To discourage excessive speech modification, we minimize

$$\begin{aligned} \mathcal{L} &= \lambda_{\text{spk}} \mathcal{L}_{\text{spk}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{energy}} \mathcal{L}_{\text{energy}}, \\ \mathcal{L}_{\text{rec}} &= \sum_{r \in \mathcal{R}} \left| \|\text{STFT}_r(x)\| - \|\text{STFT}_r(\tilde{x})\| \right|_1. \end{aligned} \quad (2)$$

The set $\mathcal{R}$ indexes short-time Fourier transform (STFT) resolutions. The energy penalty is $\mathcal{L}_{\text{energy}} = |\text{RMS}(\tilde{x}) - \text{RMS}(x)|$, where RMS denotes root-mean-square amplitude. The λ coefficients are fixed weights. Both utility terms compare pre-channel signals, so modifications removed by the channel are still penalized.

2.2. Causal spectral implementation

For 16-kHz mono speech x , let $X = \text{STFT}(x)$. At frame t , a causal predictor uses only past and current context to produce unconstrained outputs $\hat{G}_{b,t}$ for frequency band b . We parameterize the gains around unity, corresponding to no spectral modification:

$$\begin{aligned} G_{b,t} &= 1 + \alpha \tanh(\hat{G}_{b,t}), & 0 \leq \alpha < 1, \\ \tilde{X}_{j,t} &= G_{b(j),t} X_{j,t}. \end{aligned} \quad (3)$$

Here j indexes frequency bins, $b(j)$ is a fixed bin-to-band mapping, and α bounds the gain deviation from unity. Since $|\tanh(\cdot)| \leq 1$, the gains satisfy $|G_{b,t} - 1| \leq \alpha$ and $G_{b,t} \in [1 - \alpha, 1 + \alpha]$. We use a fixed $\alpha = 0.18$ in the reported experiments, allowing at most 18% relative magnitude modulation per coefficient and yielding the derived gain bounds $[0.82, 1.18]$. Zero predictor output gives the identity gain, while $\alpha < 1$ ensures positive gains and preserves coefficient phases. This constraint bounds spectral modification, not perceptual distortion; speech utility is evaluated separately.

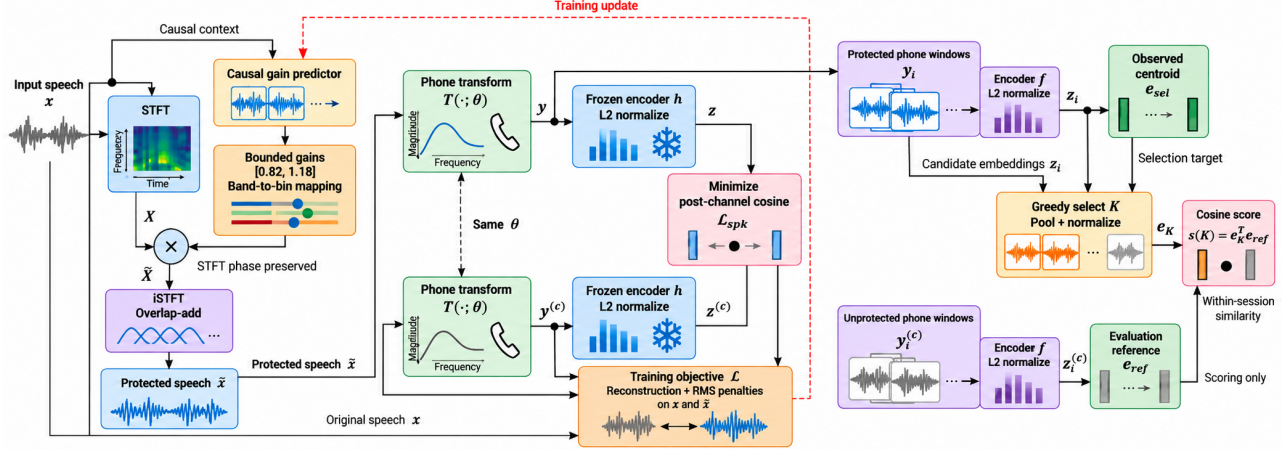


Fig. 2. CloneBlock-RT protects speech before telephone transmission. Training compares matched post-channel embeddings; captured clips support separate reference-pooling and voice-cloning evaluations.

Inverse STFT with overlap-add reconstructs $\tilde{x}$. The gains adapt to the available speech context, while inference requires only a causal predictor pass and signal reconstruction. Channel simulation and the speaker encoder are used for training, not by the deployed protector. Thus, telephone-aware optimization does not require channel estimation or iterative optimization during a call.

2.3. Reference collection and evaluation

Sessions concatenate distinct same-speaker utterances with short random silences. We extract 1-s windows at a 0.25-s hop, discard low-energy windows, and retain $N \geq 64$ candidates. Here K counts windows, not unique seconds. For the resulting channel-processed windows $y_i, y_i^{(c)}$ and an evaluation encoder f , define $z_i = \nu(f(y_i))$ and $z_i^{(c)} = \nu(f(y_i^{(c)}))$. For selected indices S_K with $|S_K| = K$,

$$\begin{aligned} e_{\text{sel}} &= \nu\left(\sum_{i=1}^N z_i\right), \quad e_{\text{ref}} = \nu\left(\sum_{i=1}^N z_i^{(c)}\right), \\ s(K) &= \nu\left(\sum_{i \in S_K} z_i\right)^T e_{\text{ref}}. \end{aligned} \quad (4)$$

Starting empty, greedy selection adds the unused window whose pooled embedding is most similar to e_{sel} . Only protected audio guides selection; e_{ref} is for scoring. This heuristic need not maximize $s(K)$. The same-session reference yields a within-session similarity score, not independent-enrollment verification or impersonation success.

For $\mathcal{K} = \{1, 2, 4, 8, 16\}$, let $\bar{s}(K)$ be the session-mean score. MeanSim@16 is $\bar{s}(16)$, AUC_K is $|\mathcal{K}|^{-1} \sum_{K \in \mathcal{K}} \bar{s}(K)$, and Slope@16-1 is $\bar{s}(16) - \bar{s}(1)$. AUC_K is an equal-weight mean, not an integrated area; Slope is an endpoint change, not absolute protection. Separate cloning tests use audio clips, not pooled embeddings. Word error rate (WER), short-time objective intelligibility (STOI), and real-time factor (RTF; processing time/audio duration) assess utility and computational cost. Reference aggregation is evaluated, not included in the training loss.

3. EXPERIMENTS

3.1. Data and telephone conditions

We use 16-kHz mono VCTK v0.92 [11] with 78/10/20 speaker-disjoint training/validation/test speakers. Each test speaker contributes five 30-s sessions (100 total), concatenating distinct utterances with short silences. We sample 1-s windows at a 0.25-s hop, uniformly retaining 96 candidates above -40 dBFS RMS per session, regenerating sessions with insufficient candidates. Seeds are fixed; windows may overlap.

Training randomizes band-limiting, resampling, AGC, and Gaussian noise. Reference and cloning use codec-free 300–3400-Hz filtering, an 8-kHz intermediate rate, -21 -dBFS AGC, and 25-dB SNR. Branches share channel settings and noise, maintaining processing states throughout each session. Utility conditions are “none” (no processing), 16-kb/s Opus [9], and 8-kHz G.711 μ -law [10]; decoded audio returns to 16 kHz.

3.2. Systems and implementation

All protection precedes transmission. Baselines comprise no protection and paper-based implementations of full offline ClearMask [4] and speaker-specific Enkidu [8], using published settings. ClearMask uses training-speaker target voices and internal 48-kHz processing before 16-kHz resampling. Enkidu uses noise level 0.4, 120-frame perturbations, and 20 enrollment utterances per speaker excluded from evaluation. Settings are frozen before testing. Frozen SpeechBrain `spkrec-ecapa-voxceleb` ECAPA-TDNN checkpoint [12] provides training/evaluation embeddings, L2-normalized for cosine scoring.

CloneBlock-RT uses two causal 64-unit GRU layers and a bounded 32-band gain head with mel spacing over 0–8 kHz. Non-centered STFT uses 320-sample Hann windows, 160-sample hops, and a 512-point FFT; interpolated gains preserve phase. Training uses Adam (10^{-3} learning rate), sixteen 4-s crops per batch, and 50,000 updates. Similarity, reconstruction, and temporal-smoothness weights are fixed at 1/10/0.1. Minimizing validation mean similarity subject to channel-free STOI ≥ 0.98 selects $\alpha = 0.18$ from $\{0.10, 0.14, 0.18, 0.22\}$; checkpoints follow the same criterion.

Table 1. Reference and XTTS cosine similarity ($\downarrow$). Mean equally weights five reference counts; reference $K = 1$ values are reconstructed from rounded summaries. Bold marks column minima.

System	Reference			XTTS output		
	$K = 1$	$K = 16$	Mean	$K = 1$	$K = 16$	$K = 64$
No defense	0.7202	0.9861	0.8830	0.70	0.88	0.92
ClearMask	0.6593	0.9326	0.8558	0.62	0.79	0.85
Enkidu	0.6433	0.9018	0.8519	0.57	0.74	0.81
CloneBlock-RT	0.6733	0.9203	0.8245	0.58	0.72	0.78

Back-propagation covers filtering, resampling, smooth AGC, and additive noise; codecs are evaluation-only.

3.3. Reference selection and cloning evaluation

Greedy observed-centroid selection uses only each condition’s captured embeddings. For $K \in \{1, 2, 4, 8, 16\}$, pooled embeddings are scored against the normalized mean of all unprotected candidate embeddings. Scores are averaged across sessions; mean over K equally weights the five counts.

Cloning uses Coqui TTS 0.22.0 with XTTS-v2 checkpoint v2.0.2 [15] for all four systems, 100 sessions, ten fixed English LibriSpeech test-clean sentences absent from references, and seeds 0/1/2 (3,000 outputs per condition and K). For $K \in \{1, 16, 64\}$, windows form file lists in greedy order, retaining overlaps. Conditioning uses `gpt_cond_len=30`, `gpt_cond_chunk_len=4`, and `max_ref_len=30` s; reference normalization and silence trimming are disabled, with default decoding. Speaker conditioning averages files; GPT conditioning uses the first 30 concatenated seconds or less. Outputs resampled from 24 to 16 kHz are scored against the unprotected session centroid without telephone processing. Averaging proceeds over generations, texts, then sessions.

3.4. Speech utility and runtime

Utility uses 100 additional VCTK test utterances and 100 LibriSpeech 2015 test-clean utterances [16] (20 speakers per corpus, five utterances each). WER uses `facebook/wav2vec2-base-960h`, greedy CTC decoding, lowercasing, punctuation removal, and whitespace normalization. STOI [17] uses `pystoi 0.4.1` against condition-matched unprotected waveforms after codec-delay compensation and common-length cropping. Both metrics report utterance means and sample standard deviations.

Streaming uses an NVIDIA A100 (FP32, batch one, 20/40-ms chunks). Timing includes analysis, gain prediction, synthesis, and GPU transfers, excluding codecs and file I/O. Ten 60-s runs follow 100 warm-up chunks, with GPU synchronization. RTF is processing time/audio duration. Initial STFT buffering is 20 ms; the predictor has no look-ahead.

4. RESULTS AND ANALYSIS

4.1. Similarity under multi-window reference selection

The reference columns in Table 1 show reductions for every defense. CloneBlock-RT lowers mean similarity by 0.0585 versus no defense and 0.0274 versus Enkidu. Enkidu nevertheless performs best at both endpoints, beating CloneBlock-RT by 0.0185 at $K = 16$. Thus, the average advantage is not uniform across window counts.

Table 2. Utility over 200 utterances: mean $\pm$ standard deviation. WER and its deviation are percentages; lower WER and higher STOI are preferred.

Condition	System	WER (%)	STOI
none	No defense	52.84 $\pm$ 24.88	1.0000 $\pm$ 0.0000
none	CloneBlock-RT	52.14 $\pm$ 23.81	0.9887 $\pm$ 0.0047
Opus 16 kb/s	No defense	54.27 $\pm$ 25.53	1.0000 $\pm$ 0.0000
Opus 16 kb/s	CloneBlock-RT	54.68 $\pm$ 25.89	0.9242 $\pm$ 0.0410
G.711	No defense	53.76 $\pm$ 24.52	1.0000 $\pm$ 0.0000
G.711	CloneBlock-RT	53.09 $\pm$ 24.47	0.9823 $\pm$ 0.0119

Endpoint increases are 0.2659 without protection and 0.2470 with CloneBlock-RT; this difference neither establishes stronger protection independently of starting scores nor guarantees monotonic behavior under centroid-based selection.

4.2. Speaker similarity of cloned speech

All defenses lower XTTS similarity (Table 1). ClearMask reduces scores by 0.07–0.09 and Enkidu by 0.11–0.14, compared with 0.12/0.16/0.14 for CloneBlock-RT at $K = 1/16/64$. Enkidu leads at $K = 1$ (0.57 versus 0.58), while CloneBlock-RT is lower by 0.02/0.03 at $K = 16/64$ and by 0.07 relative to ClearMask at both counts. Rankings therefore depend on reference count and differ from pooled-reference endpoints. Similarity increases with K for every system; overlapping windows and conditioning limits preclude interpreting K as unique speech duration.

4.3. Utility and computational cost

WER changes are -0.70 , $+0.41$, and -0.67 percentage points for none, Opus, and G.711 (Table 2). These describe incremental effects on an already high-error pipeline, not recognition equivalence. Utterance deviations are not confidence intervals for paired differences. STOI declines most under Opus (0.9242); unity baselines follow from condition-matched references and do not imply distortion-free transmission.

For 20/40-ms chunks, CPU RTF is 0.036/0.025 (0.72/1.00 ms per chunk), and GPU RTF is 0.019/0.013 (0.38/0.52 ms). All configurations provide real-time throughput, without establishing worst-case processing time or end-to-end call latency.

5. CONCLUSION

We presented CloneBlock-RT, a single-pass causal predictor of bounded, frame-dependent STFT gains, trained with a speaker-similarity objective evaluated after matched telephone processing. Across tested reference counts, it reduces mean reference similarity from 0.8830 to 0.8245 and XTTS similarity by 0.12–0.16, although relative rankings vary. Mean WER changes stay within 0.70 percentage points, while STOI indicates codec-dependent costs. Reported A100 timings support real-time throughput. These findings support similarity reduction under multi-reference collection. Future work will assess impersonation risk through independent-enrollment verification, transfer across speaker encoders and cloning backends, and ablations of channel sampling and adaptive gains.

6. REFERENCES

- [1] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath, “Voicecraft: Zero-shot speech editing and text-to-speech in the wild,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 12442–12462, Association for Computational Linguistics.
- [2] Edresson Casanova, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A. Ponti, “YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone,” in *Proceedings of the 39th International Conference on Machine Learning*. 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 2709–2720, PMLR.
- [3] Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, and Qiben Yan, “Vsmask: Defending against voice synthesis attack via real-time predictive perturbation,” in *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2023, pp. 239–250.
- [4] Yuanda Wang, Bocheng Chen, Hanqing Guo, Guangjing Wang, Weikang Ding, and Qiben Yan, “Clearmask: Noise-free and naturalness-preserving protection against voice deepfake attacks,” in *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIA CCS ’25)*, Hanoi, Vietnam, August 2025, pp. 696–709, ACM.
- [5] Renyuan Li, Zhibo Liang, Haichuan Zhang, Tianyu Shi, Zhiyuan Cheng, Jia Shi, Carl Yang, and Mingjie Tang, “Cloneshield: A framework for universal perturbation against zero-shot voice cloning,” *arXiv preprint arXiv:2505.19119*, 2025.
- [6] Zhisheng Zhang, Derui Wang, Qianyi Yang, Pengyang Huang, Junhan Pu, Yuxin Cao, Kai Ye, Jie Hao, and Yixian Yang, “Safe-speech: Robust and universal voice protection against malicious speech synthesis,” in *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*, Seattle, WA, USA, August 2025.
- [7] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang, “Antifake: Using adversarial audio to prevent unauthorized speech synthesis,” in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 460–474.
- [8] Zhou Feng, Jiahao Chen, Chunyi Zhou, Yuwen Pu, Qingming Li, Tianyu Du, and Shouling Ji, “Enkidu: Universal frequential perturbation for real-time audio privacy protection against voice deepfakes,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11638–11647.
- [9] Jean-Marc Valin, Koen Vos, Timothy B. Terriberry, et al., “RFC 6716: Definition of the opus audio codec,” RFC Editor / IETF, September 2012.
- [10] “ITU-T recommendation G.711: Pulse code modulation (pcm) of voice frequencies,” Recommendation G.711, International Telecommunication Union (ITU-T), November 1988.
- [11] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive*: (<http://web.ku.edu/~idea/readings/rainbow.htm>), 2019.
- [12] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [13] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Gökmar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber, “XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model,” in *Interspeech 2024*, 2024, pp. 4978–4982.
- [14] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, “Synthesizing robust adversarial examples,” in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 284–293, PMLR.
- [15] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Gökmar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber, “XTTS: A massively multilingual zero-shot text-to-speech model,” in *Proc. Interspeech*, 2024, pp. 4978–4982.
- [16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210, IEEE.
- [17] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.