LM-AdvVC: Localized Multi-Granularity Targeted Adversarial Attacks on Voice Conversion

1st Xinning Song
Tongji University
Shanghai, China
xinningsong@tongji.edu.cn

2nd Rui Wang
IFLYTEK Research
Shanghai, China
rwang@tongji.edu.cn

3rd Liwei Zheng
Tongji University
Shanghai, China
2432010@tongji.edu.cn

4th Zhihua Wei*
Tongji University
Shanghai, China
zhihua_wei@tongji.edu.cn

***Abstract*—Despite recent progress in generating natural and personalized speech, voice conversion (VC) systems that aim to modify speaker identity while preserving linguistic content remain vulnerable to degraded inputs, where even minor adversarial perturbations cause severe performance drops. Prior studies mainly focused on perturbations that affect speaker identity in classic encoder-decoder models (e.g., AdaIN-VC, AutoVC), while the impact of linguistic-targeted perturbations, especially on modern neural audio codecs, remains unexplored. To address this gap, we propose LM-AdvVC, a two-stage localized adversarial attack framework that applies speaker-guided perturbations directly to content embeddings. Firstly, a critical-segment localization with phoneme- and word-level analysis adaptively identifies and perturbs vulnerable regions. Secondly, the waveform perturbations are refined using CMA-ES optimization under a teacher-forcing cross-entropy objective, yielding transferable adversarial attacks across gray-box scenarios. Experimental results demonstrate that LM-AdvVC successfully manipulates linguistic content under both white and gray-box settings and reveal the content vulnerability of VC systems. To ensure reproducibility, the code and supplementary materials are available at https://github.com/windforestfiremountain/LM-AdvVC.**

***Index Terms*—Voice Conversion; Adversarial Attack; Multi-level Local Perturbation; Evolutionary Strategy; Gray Box Test**

I. Introduction

Voice conversion (VC) [1] is a style-transferring task that aims to preserve linguistic information while altering speaker's identity into specified tones of the target. Driven by the remarkable progress in neural vocoders [2], [3], generative models (e.g., GANs [4], VAEs [5], and Diffusion [6], [7] models), as well as self-supervised learning (SSL) models [8], [9], VC achieves unprecedented realism and widely applies in various domains, including audio editing [10], dysarthric speech recovery [11], privacy preservation [12], and data augmentation [13]. This widespread success is largely predicated on clean and studio-quality inputs but may suffer from audio distortion under more realistic *in the wild* conditions.

To investigate the robustness of VC systems, researchers employ adversarial attacks to expose the vulnerabilities of neural networks. Such attacks introduce imperceptible yet carefully crafted perturbations to the input and can reliably manipulate the model into producing attacker-specified outputs. Current researchers focus on time-frequency perturbations to promote the effectiveness, imperceptibility, and transferability. Specifically, some methods optimize the perturbation directions in latent space using the gradients of target models to obtain white-box attacks [14]–[17]. Some researchers enhance auditory imperceptibility through techniques such as SNR constraints, mean subtraction, and psychoacoustic model [18], [19]. To address challenges of the model transferability against VC systems, approaches such as natural evolution strategies (NES) and ensembled encoders have also been explored [20], [21]. However, these approaches still suffer two essential challenges: (1) a narrow focus on speaker identity while neglecting linguistic attributes, and (2) practically infeasible optimization over the large linguistic-information search space.

To address these limitations, we propose a two-stage localized adversarial attack framework, called LM-AdvVC. In the first stage, a critical-segment localization strategy exploits the Encoder–RVQ–Decoder architecture to narrow the search space, followed by a speaker-guided multi-granularity analysis at both phoneme and word levels to manipulate vulnerable segments in substitute models. In the second stage, the resulting perturbation is further refined by the CMA-ES algorithm, which queries the teacher-forcing cross-entropy loss to adjust the Gaussian noise distribution, and produces the final adversarial example.

The contributions are summarized as follows.

- We present LM-AdvVC, a two-stage localized adversarial attack framework that first localizes and perturbs critical segments via multi-granularity analysis and then refines with CMA-ES under a teacher-forcing cross-entropy objective.
- We propose a critical-segment localization method that leverages CTC alignments to locate and adaptively expand the most vulnerable word-level region for targeted perturbation.
- We construct and release a 1,367-sentence evaluation benchmark derived from the VCTK corpus. Mispronunciations are systematically generated by a pronunciation dictionary and edit-distance rules, designed to evaluate robustness under pause, and digit scenarios.

*Corresponding author.

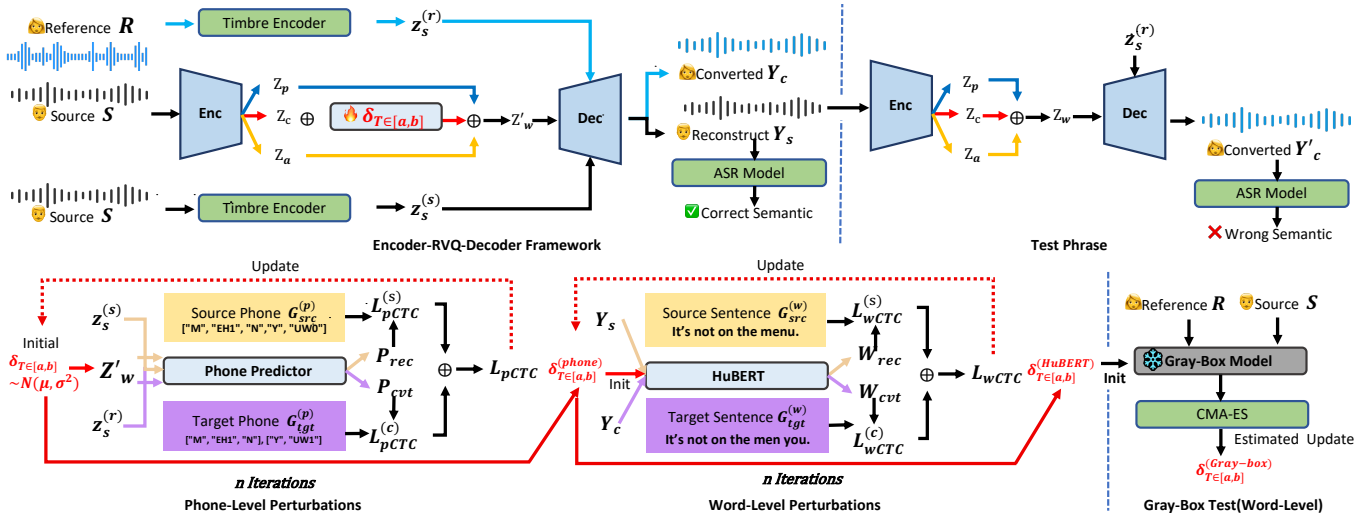

Fig. 1: Overview of **LM-AdvVC**. **Top:** Encoder–RVQ–Decoder voice-conversion framework; a successful attack yields a perturbed source reconstruction Y_s that preserves linguistic information while the converted output Y_c is misrecognized. **Bottom:** Two-stage iterative attack: a localized perturbation $\delta_T \in [a, b]$ is initialized from a Gaussian and first optimized with phone-level CTC loss L_{pCTC}. Updates are then refined with word-level CTC loss L_{wCTC} using HuBERT (initialized from the phone-level delta). Finally, a gray-box word-level attack is tuned with CMA-ES via teacher-forced cross-entropy queries.

II. Related Works

A. Voice Conversion and Neural Audio Codecs

Voice Conversion (VC) [22] is typically formulated as a regression problem that maps source speech features to target speaker attributes while preserving linguistic content. Over the past decade, VC methods have evolved from statistical learning to deep generative and diffusion-based paradigms. Early approaches, such as GMM- and VQ-based methods, relied on parallel data and suffered from limited scalability. To overcome this, deep generative models enabled non-parallel training: StarGAN-VC [4] leveraged cycle-consistency for many-to-many conversion, while AutoVC [23] introduced an information bottleneck to disentangle speaker and content representations for zero-shot transfer. More recently, diffusion-based models [6], [7], [24] have further improved audio fidelity by modeling speech generation as a stochastic denoising process, achieving state-of-the-art quality.

In parallel, Neural Audio Codecs have emerged as a powerful paradigm for efficient and high-quality speech representation learning. Methods such as SoundStream [25] and EnCodec [26] employ encoder–decoder architectures with residual vector quantization (RVQ) to compress speech into discrete latent tokens, though these representations often entangle linguistic and paralinguistic information. To address this limitation, FACodec, built upon the NaturalSpeech3 [27] framework, introduces factorized vector quantization (FVQ) to decompose speech into multiple independent latent subspaces, enabling improved controllability and interpretability. This factorized design has made FACodec a strong backbone for codec-based VC systems **FACodec-VC**, particularly in zero-shot settings. However, despite its effectiveness, the robustness and security of such explicitly disentangled representations remain underexplored, motivating our focus on analyzing their vulnerability under adversarial perturbations.

B. Adversarial Attacks on Voice Conversion

While adversarial attacks on ASR and ASV have been extensively studied, those targeting Voice Conversion (VC) remain relatively underexplored. Early efforts in white-box settings primarily relied on gradient-based optimization in the frequency domain to disrupt speaker representations, including feedback-driven strategies [14] and iterative methods such as I-FGSM [17]. To improve perceptual quality and efficiency, later works incorporated generative priors, leveraging GANs [28] and diffusion models [29] to better align adversarial perturbations with natural speech distributions. These approaches were further extended to time-domain settings, where psychoacoustic modeling and dynamic masking techniques were introduced to balance imperceptibility and attack effectiveness across real-time [15], [16], universal [30], waveform-level scenarios [18], [31].

In black-box settings, gradient-free methods such as ensemble-based attacks [20] and Natural Evolution Strategies (NES) [21] have been explored to estimate gradients and improve transferability. More specialized studies, such as those in singing voice conversion [32], have begun to investigate semantic-level perturbations targeting lyrics. Despite these advances, existing VC attacks predominantly focus on global timbre manipulation and show limited effectiveness against modern codec-based VC systems with discrete latent representations. In contrast, this work shifts the focus to linguistic content in codec-driven VC models, proposing a localized,

content-oriented attack framework to systematically reveal vulnerabilities in factorized speech representations.

III. Methodology

We aim to generate localized lingustic adversarial examples for voice conversion systems. As shown in Fig. 1, a successful attack must ensure that the reconstructed source audio Y_s retains its original semantics, while the converted audio Y_c with perturbed but speaker-guided linguistic content, is misrecognized by ASR systems. To this end, we propose LM-AdvVC, a two-stage attack framework. The localized perturbation $\delta_{T\in[a,b]}$ is optimized in three sequential stages: first at the phone level using CTC loss (L_{pCTC}) in a white-box setting; next at the word level with HuBERT (L_{wCTC}); and finally searched via CMA-ES [1] in gray-box settings.

A. Problem Definition

Voice Conversion System. We consider a codec-based voice conversion framework (FACodec-VC) [27] built on a residual vector quantization (RVQ) encoder–decoder architecture[2]. As illustrated in Fig. 1, given a source utterance S and a reference utterance R, the encoder decomposes S into three disentangled latent codes: prosody z_p, content z_c, and acoustic-detail z_a, while a timbre encoder extracts speaker embeddings from R and S:

$$(z_p, z_c, z_a) = \mathrm{Enc}(S), \tag{1}$$

$$z_s^{(r)} = \mathrm{TimbreEnc}(R), \quad z_s^{(s)} = \mathrm{TimbreEnc}(S). \tag{2}$$

To enable adversarial manipulation, we inject a time-localized perturbation $\delta_{T\in[a,b]}$ into the content latent space z_c:

$$z_c' = z_c + \delta_{T\in[a,b]}. \tag{3}$$

Subsequently, the decoder generation process strictly follows the inverse transformation flow. Specifically, the perturbed content features z_c' are first re-aggregated with prosody z_p and acoustic details z_a, and then modulated by the AdaIN module conditioned on source timbre $z_s^{(s)}$ and target timbre $z_s^{(r)}$, respectively. Finally, the generator reconstructs the time-domain waveform:

$$Y_s = \mathrm{Dec}(z_p, z_c', z_a, z_s^{(s)}) \approx \mathrm{Gen}(\mathrm{AdaIN}(z_c' \oplus z_p \oplus z_a, z_s^{(s)})), \tag{4}$$

$$Y_c = \mathrm{Dec}(z_p, z_c', z_a, z_s^{(r)}) \approx \mathrm{Gen}(\mathrm{AdaIN}(z_c' \oplus z_p \oplus z_a, z_s^{(r)})). \tag{5}$$

Attack Goal. Based on the above scenario, formally, given source S and reference R, the attacker aims to find a subtle localized perturbation δ to be superimposed on the content representation z_c. Given a malicious target text or phoneme sequence $G_{tgt}^{(x)}$ ($x \in \{word, phone\}$) preset by the attacker, and the original source text or phoneme sequence $G_{src}^{(x)}$, a successful adversarial attack must simultaneously satisfy the following two constraints:

(1) Reconstruction Invariance: The adversarial perturbation should not disrupt the original linguistic content of the source speech, ensuring the manipulation is imperceptible to human ears. The reconstructed speech Y_s must remain semantically faithful to the source text $G_{src}^{(w)}$, i.e., $\mathrm{ASR}(Y_s) = G_{src}^{(w)}$.

(2) Target Misleading: The converted speech Y_c should be misrecognized by the target ASR system as the attacker-specified malicious target text $G_{tgt}^{(w)}$, i.e., $\mathrm{ASR}(Y_c) = G_{tgt}^{(w)}$.

Threat Model. We consider two typical adversarial scenarios based on the attacker's knowledge of the target model:

(1) White-box Attack: In this setting, the attacker has full access to the target VC model parameters (e.g., FACodec encoder) and possesses gradient access to the evaluated ASR system (e.g., HuBERT[3]). The attacker calculates the gradient of the CTC loss function with respect to the perturbation δ, denoted as $\nabla_\delta \mathcal{L}_{CTC}$, via backpropagation to directly update the perturbation:

$$\delta \leftarrow \delta - \eta \nabla_\delta \mathcal{L}(\delta). \tag{6}$$

(2) Gray-box Attack: In this setting, we define the threat as 'gray-box' to distinguish it from a fully opaque black-box scenario. Here, the attacker cannot access the internal gradients of the target model or the ASR system. The attacker only knows the interface of the target ASR system (e.g., Whisper Tiny[4]). We utilize a gradient-free optimization algorithm, such as CMA-ES, to query the ASR system's output text and iteratively refine the perturbation distribution until the constraints of reconstruction invariance and target misleading are met:

$$\delta^* \approx H\big(\mathrm{F}(\mathrm{Whisper}(Y_c), G_{tgt}^{(w)}), \mu^{(0)}, \sigma^{(0)}\big), \tag{7}$$

where $\mu^{(0)}$ and $\sigma^{(0)}$ denote the mean and variance of the Gaussian initialization noise, and F represents the fitness function.

B. Localization Strategy

To precisely localize perturbation regions, we propose a multi-step local mask generation strategy. Let the ground-truth source and target phoneme sequences be $G_{src} = (g_1^{src}, \ldots, g_M^{src})$ and $G_{tgt} = (g_1^{tgt}, \ldots, g_M^{tgt})$, respectively. We employ a Conformer-based phone predictor, trained on the VCTK corpus with CMU Dict phoneme annotations[5], to obtain the predicted sequence $P_{src} = (p_1^{src}, \ldots, p_M^{src})$ from input waveform $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$. Here, P_{src} records phoneme identities, while $T_{src} = (t_1^{src}, \ldots, t_M^{src})$ denotes the end-frame indices along the temporal axis.

We first align G_{src} and G_{tgt} to identify the minimal mismatched region. This region is expanded to the nearest padding boundaries to ensure word-level consistency. By subsequence matching, the expanded source segment is mapped to P_{src}, yielding an index range $[k_L, k_R]$. Using T_{src}, this range is translated into a temporal interval $[s, e]$, where $s = \max(0, t_{k_L-1}^{src})$ and $e = t_{k_R}^{src}$. To account for alignment uncertainty, the interval is extended to $[s', e']$ by including

[1] https://github.com/CMA-ES/pycma

[2] https://huggingface.co/spaces/amphion/naturalspeech3_facodec/tree/main

[3] https://huggingface.co/facebook/hubert-base-ls960

[4] https://huggingface.co/openai/whisper-tiny

[5] https://github.com/Kyubyong/g2p

Algorithm 1 Gradient Estimate for Word-Level Attack

Input: latent content z_c, other latents (z_p, z_a, z_s); Decoder $\mathrm{Dec}(\cdot)$; Whisper Tiny query $\mathrm{Whisper}(\cdot)$;
Output: adversarial converted audio Y_c
Initialize local perturbation $\delta^{(0)}$ (from Stage I), mask M (localized perturb except in $[a, b]$); Initial mean $m^{(0)} = \delta^{(0)}$, $\sigma^{(0)}$, population λ, generations G, weights $\{w_i\}_{i=1}^{\mu}$

for $g = 0, \ldots, G-1$ **do**
  Sample λ candidates:

$$x_k^{(g)} \sim \mathcal{N}\big(m^{(g)},\ {\sigma^{(g)}}^2 I\big), \quad k = 1, \ldots, \lambda$$

  for $k = 1, \ldots, \lambda$ **do**
    $\tilde{\delta}_k \leftarrow M \odot x_k^{(g)}$ (only frames in $[a, b]$ modified)
    $Y_{c,k} \leftarrow \mathrm{Dec}(z_p,\ z_c + \tilde{\delta}_k,\ z_a,\ z_s)$
    $\hat{y}_k \leftarrow \mathrm{Whisper}(Y_{c,k})$
    $f_k \leftarrow \mathrm{F}(\hat{y}_k, G_{tgt}^{(w)})$ (fitness function; lower is better)
  end for
  Sort $\{x_k^{(g)}\}$ by ascending f_k to obtain $x_{1:\lambda}^{(g)}$
  Aggregate elites to update the mean:

$$m^{(g+1)} \leftarrow \sum_{i=1}^{\mu} w_i\, x_{i:\lambda}^{(g)}$$

  Optionally update covariance $\sigma^{(g+1)}$
  Project $m^{(g+1)}$ to satisfy $\|M \odot m\|_\infty \le \epsilon$
end for
$\delta^* \leftarrow m^{(G)}$
$Y_c \leftarrow \mathrm{Dec}(z_p,\ z_c + (M \odot \delta^*),\ z_a,\ z_s)$
return Y_c

adjacent blank frames or by adding a fixed margin δ_{margin}. The resulting region is $R^{src} = [s', e']$.

If a global adversarial attack succeeds, we can additionally derive R^{adv} from (P^{adv}, T^{adv}). The final perturbation mask is then obtained by union:

$$M = [a, b] = R^{src} \cup R^{adv}. \tag{8}$$

C. LM-AdvVC

1) Stage I: Multi-Granularity Analysis: Based on the localized perturbation mask M, we propose a multi-granularity targeted perturbation method that jointly operates at the phoneme and word levels.

Phoneme-level attack: In this stage, we enforce phoneme sequence alignment between the adversarially reconstructed speech Y_s and the source phoneme sequence $G_{src}^{(p)}$, while simultaneously pushing the adversarially converted speech Y_c toward the target $G_{tgt}^{(p)}$. We employ a pre-trained phone predictor Φ that maps waveforms into phoneme prediction sequences:

$$P_{rec} = \Phi(Y_s), \quad P_{cvt} = \Phi(Y_c). \tag{9}$$

The loss is defined as a weighted combination of two CTC terms:

$$L_{pCTC}(\delta) = (1-\lambda_p)\mathrm{CTC}(P_{rec}, G_{src}^{(p)}) + \lambda_p \mathrm{CTC}(P_{cvt}, G_{tgt}^{(p)}), \tag{10}$$

where λ_p balances source preservation and target enforcement.

Word-level attack: Our word-level attack builds upon the phoneme-level stage. The perturbation is initialized from the converged phoneme-level perturbation $\delta_{T\in[a,b]}^{(phone)}$ and then generate Y_s and Y_c by employing FACodec VC. We adopt the HuBERT Ψ, which predicts word sequences from speech:

$$W_{rec} = \Psi(Y_s), \quad W_{cvt} = \Psi(Y_c). \tag{11}$$

Following the same formulation, the word-level loss is:

$$L_{wCTC}(\delta) = (1-\lambda_w)\mathrm{CTC}(W_{rec}, G_{src}^{(w)}) + \lambda_w \mathrm{CTC}(W_{cvt}, G_{tgt}^{(w)}), \tag{12}$$

where λ_w controls the trade-off between preserving the source and enforcing the adversarial target at the word level.

By combining these two granularities(phone and word-level), our approach achieves stronger and more reliable linguistic adversarial attacks in codec-based VC systems.

2) Stage II: Gray-Box Analysis: Stage I produces a localized $\delta_{T\in[a,b]}^{(\mathrm{HuBERT})}$ as the initial $\delta^{(0)}$ in Stage-II(gray-box). Covariance matrix adaptation evolution strategy (CMA-ES) is employed to iteratively sampling candidate solutions x_k from a parameterized Gaussian distribution, and evaluated via a fitness function F. It measures the alignment between the predicted adversarial output transcription $\hat{y_k}$ and the target $G_{tgt}^{(w)}$, and records as f_k. The candidates are sorted by f_k and estimated δ^* is approximated by weighted recombination of the top μ samples

$$m^{(g+1)} = \sum_{i=1}^{\mu} w_i\, x_{i:\lambda}^{(g)} \tag{13}$$

After G generations, δ^* is updated by $m^{(G)}$, and thereby Y_c is generated by

$$Y_c = \mathrm{Dec}(z_p,\ z_c + (M \odot \delta^*),\ z_a,\ z_s) \tag{14}$$

An early-stop condition terminates evolution once Y_c recognizes as the expected target $G_{tgt}^{(w)}$, as detailedly illustrated in Algorithm 1.

IV. Experiments and Results

A. Experimental Setup

Dataset and Lexicon Construction: We conduct experiments on the VCTK corpus [33]. To ensure precise phoneme-to-word alignment, we implement a standardized data processing pipeline as shown in Fig. 2. We integrate the CMU Dict and EnglishDict[6] to generate initial phoneme sequences. Through strict deduplication and lexicographical sorting, we distill approximately 44k raw utterances into a compact pronunciation lexicon V_w containing 5,858 high-quality entries. Crucially, V_w preserves complex "form-sound" mappings,

[6]https://github.com/kajweb/dict

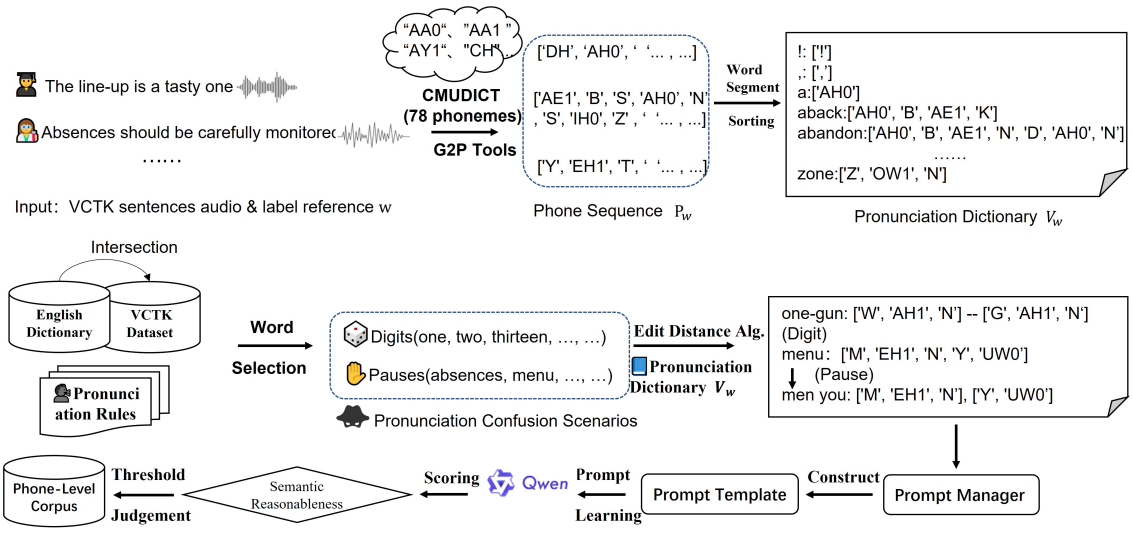

Fig. 2: The automated pipeline for pronunciation dictionary construction and adversarial benchmark generation. The process distills VCTK data into a standardized lexicon V_w and generates confusion scenarios (Digits, Pauses) via Levenshtein edit distance, filtered by a triple-semantic verification mechanism).

explicitly recording polyphones (e.g., *close* as /z/ vs. /s/) and homophones (e.g., *threw*/*through*) to support fine-grained adversarial generation.

To evaluate robustness against specific linguistic perturbations, we propose an automated generation pipeline (Fig. 2). We employ Levenshtein edit distance (threshold $\tau = 3$) to generate candidate adversarial pairs. To guarantee linguistic plausibility, we apply a **triple-semantic filtering mechanism**: (1) *WordNet* validates lexical existence; (2) *SBERT* ensures sentence-level semantic similarity; and (3) *Qwen LLM* scores fluency via few-shot prompting, filtering out low-quality samples. From the filtered pool, we construct a 1,367-sentence evaluation benchmark covering two core error-prone scenarios:

- **Digit Confusion:** We target numeric robustness by pairing digits with phonetically similar words ($\tau < 3$). A typical example is replacing *one* (/W AH1 N/) with *gun* (/G AH1 N/), where the edit distance is 1.
- **Pause Anomalies:** We induce artificial pauses by splitting words into valid sub-words to test prosodic coherence. For instance, *menu* (/M EH1 N Y UW0/) is split into *men* and *you*. The combined phonemes remain identical (edit distance 0) but introduce a boundary interference.

Training and Hyperparameters: The Phone Predictor Φ is trained with a Conformer backbone on 43,567 utterances, each represented as latent features $Z_c \in \mathbb{R}^{B\times C\times T}$. Another 500 utterances are reserved for testing. Training uses the Cosine Annealing scheduler with an initial learning rate of 10^{-4}, minimum learning rate cycle length T_{max} 50, and warm-up $2,000$ steps. The model achieves a phoneme error rate (PER) of 6.72% on the test set. For the targeted perturbation, we set the maximum perturbation bound $\epsilon = 0.5$, $\alpha = 0.01$, and maximum iteration 1000. In the gray-box setting, CMA-ES is employed with a population size of 24 and a maximum of 300 generations.

Baseline Methods. We evaluate LM-AdvVC against five baselines covering white-box strategies such as noise injection and synthesis, as well as gray-box gradient-free optimization.

- **Global-Gaussian**: This naive baseline injects equal-energy Gaussian noise $\delta^{(G)}$ across the full temporal range of Z_c without masking

$$z'c = z_c + \delta^{(G)}, \quad \delta^{(G)} \sim \mathcal{N}(0, \sigma^2\mathbf{I}), \quad \text{s.t. } |\delta^{(G)}|\infty \leq \epsilon \tag{15}$$

  It serves as a lower bound to assess the disruptiveness of indiscriminate energy expansion.
- **Local-Gaussian** [34]: To evaluate the contribution of our localization strategy, this method injects random noise only within the masked regions.

$$z'_c = z_c + M \odot \delta^{(L)}, \quad \delta^{(L)} \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{16}$$

- **FS2-Synth (FastSpeech2 Synthesis)** [35]: This non-adversarial generative baseline directly synthesizes the target segment from text using FastSpeech2.

$$Y_{synth} = \mathrm{TTS}_{FS2}(G^{(p)}_{tgt}, Z^{(r)}_s). \tag{17}$$

- **PGD-Phone** (Ours - Stage I only): This variant represents the **ablation study** of our framework, executing only the first-stage phoneme-level optimization using the phoneme predictor Φ.
- **NES-Whisper** [36]: For the gray-box scenario, we employ Natural Evolution Strategies (NES) to estimate pseudo-gradients by querying the target ASR system Whisper Tiny:

$$\nabla_\delta \mathbb{E}[F] \approx \frac{1}{\lambda\sigma}\sum_{i=1}^{\lambda} F(z_c + M \odot (\delta + \sigma\epsilon_i)) \cdot \epsilon_i \tag{18}$$

Evaluation Metrics: We evaluate attacks from three points.

- **Attack effectiveness**: quantified by attack word success rate (AWSR), defined as the proportion of target words

successfully perturbed in VC outputs. We further report pAWSR and dAWSR as pause and digit variants, respectively.

- **Content evaluation**: quantified by reconstruct word error rate (rWER) and VC-specific word error rate (VC-WER). rWER compares ASR transcriptions between reconstructed audio and the source G_{src}, while VC-WER compares transcriptions between converted audio and the target $G_{\text{tgt}}^{(w)}$, reflecting recognition consistency before and after attack.
- **Perceptual quality**: quantified by Reconstruct PESQ (rPESQ) and Reconstruct STOI (rSTOI), which capture speech naturalness and intelligibility. These metrics also indicate that perturbations remain largely imperceptible after reconstruction, indirectly demonstrating the stealthiness of the attack.

B. Results and Discussion

1) Main Results: In the white-box evaluation (Table I), PGD-HuBERT achieves the highest attack effectiveness (68.03% AWSR) with moderate perceptual quality (rPESQ = 2.44). Interestingly, its rWER (3.42%) is even lower than Baseline (4.97%), partly due to pronunciation variations in VCTK speakers(*e.g.*, "Monitored" pronounced as "Monited"). In contrast, FS2-Synth [35], Local-Gaussian, and PGD-Phone all show limited AWSR (<17%) and similar VC-WER to Baseline. FS2-Synth fails mainly in pause cases, where small phoneme distances yield low PESQ and are easily collapsed by VC into untargeted outputs. PGD-Phone, though achieving only partial success, is necessary as it offers more targeted and interpretable optimization than Local-Gaussian, while also narrowing the search space for faster PGD-HuBERT convergence.

TABLE I: Performance of different methods under white-box scenarios. Metrics include HuBERT rWER, rPESQ, rSTOI, AWSR, and VC-WER. FS2-Synth refers to FastSpeech2 Synthesis. PGD-Phone and PGD-HuBERT represent the phone-level and word-level attacks.

Method	rWER ↓	rPESQ ↑	rSTOI ↑	AWSR ↑	VC-WER ↓
Baseline	4.97%	4.64	1.00	0.00%	21.52%
FS2-Synth	16.61%	1.30	0.16	16.97%	21.14%
Local-Gaussian	5.99%	2.85	0.87	11.63%	21.49%
PGD-Phone	7.55%	2.50	0.86	9.95%	22.26%
PGD-HuBERT	3.42%	2.44	0.84	68.03%	9.83%

Table II summarizes the results of gray-box scenarios. The Baseline yields 0% AWSR, confirming natural speech does not trigger targeted errors. FS2-Synth (16.67% AWSR) and NES-Whisper (6.25% AWSR) show only limited success with degraded quality and higher VC-WER. In contrast, CMA-ES-Whisper outperforms all others, attaining 60.42% AWSR and reducing VC-WER to 11.70% with competitive perceptual quality.

2) Analysis: Table III reports the results on the randomly selected digit (39 utterances) and pause (57 utterances) categories. CMA-ES-Whisper achieves consistently high attack success rates in both digit and pause categories, demonstrating strong robustness. In pause cases, the WER is higher because word-boundary insertions or deletions typically yield an edit distance of two, whereas digit substitutions usually involve only an edit distance of one. Consequently, NES-Whisper can occasionally succeed in pause scenarios but completely fails in digit cases.

TABLE II: Gray-box Performances of different methods. Metrics include Whisper-Tiny rWER, rPESQ, rSTOI, AWSR, and VC-WER.

Method	rWER ↓	rPESQ ↑	rSTOI ↑	AWSR ↑	VC-WER ↓
Baseline	9.89%	4.64	1.00	0.00%	26.49%
FS2-Synth	16.90%	1.26	0.18	16.67%	20.41%
NES-Whisper	14.83%	2.42	0.85	6.25%	32.06%
CMA-ES-Whisper	10.87%	2.62	0.88	60.42%	11.70%

TABLE III: Performance of different categories under gray-box scenarios. Metrics include pAWSR,dAWSR,VC-pWER, and VC-dWER.

Method	pAWSR ↑	dAWSR ↑	VC-pWER ↓	VC-dWER ↓
NES-Whisper	10.53%	0.00%	38.36%	23.35%
CMA-ES-Whisper	66.68%	51.28%	11.97%	11.28%

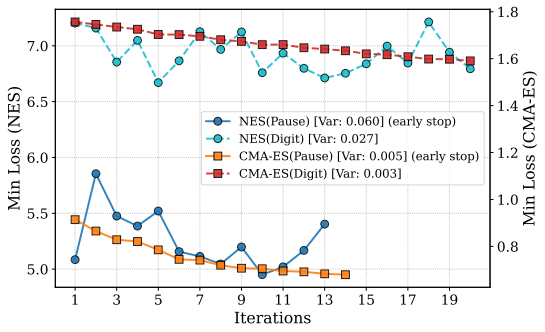

Fig. 3: Case Study under digit and pause categories.

To further understand why CMA-ES outperforms NES [37] in gray-box setting, we analyze representative samples from digit and pause categories. Fig.3 shows CMA-ES achieves a smooth descent with low variance, whereas NES oscillates and often fails to converge. Benefiting from covariance adaptation and fitness updates, CMA-ES steadily reduces loss, explaining its superior attack success and lower WER in Table III.

V. Conclusion

In this work, we proposed LM-AdvVC, a localized, multi-level targeted adversarial attack to probe vulnerabilities in speaker-guided, codec-based voice conversion systems. Validated on FACodec-VC under white-box and gray-box settings, our method can induce specific content errors (e.g., pauses, digits) while maintaining imperceptible audio quality, highlighting critical security concerns.

Future work will address the instability of phoneme-level attacks in VQ-based representations and explore localized

spectro-temporal perturbations, aiming to enhance imperceptibility and optimization robustness under both black-box and cross-dialect scenarios.

VI. Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (Nos. 62376199, 62576249, 62576247) and the Shanghai Municipal Education Commission (No. 24CGA20).

References

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[2] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[3] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[4] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[5] T. V. Ho and M. Akagi, "Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder," 2020.

[6] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation," in *Proc. Interspeech 2023*, 2023, pp. 2283–2287.

[7] F. Huang, K. Zeng, and W. Zhu, "Diffvc+: improving diffusion-based voice conversion for speaker anonymization," in *Interspeech*, vol. 2024, 2024, pp. 4453–4457.

[8] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc. Interspeech 2023*, 2023, pp. 2053–2057.

[9] S. Shan, Y. Li, A. Banerjee, and J. B. Oliva, "Phoneme hallucinator: One-shot voice conversion via set expansion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 14 910–14 918.

[10] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian *et al.*, "Audit: Audio editing by following instructions with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 71 340–71 357, 2023.

[11] H. Wang, T. Thebaud, J. Villalba, M. Sydnor, B. Lammers, N. Dehak, and L. Moro-Velazquez, "Duta-vc: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model," in *Proc. Interspeech 2023*, 2023, pp. 1548–1552.

[12] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, "The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[13] M. K. Singh, N. Takahashi, and N. Onoe, "Iteratively improving speech recognition and voice conversion," in *Proc. Interspeech 2023*, 2023, pp. 206–210.

[14] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 552–559.

[15] Z. Liu, Y. Zhang, and C. Miao, "Protecting your voice from speech synthesis attacks," in *Proceedings of the 39th Annual Computer Security Applications Conference*, 2023, pp. 394–408.

[16] Y. Wang, H. Guo, G. Wang, B. Chen, and Q. 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.

[17] S. Chen, L. Chen, J. Zhang, K. Lee, Z. Ling, and L. Dai, "Adversarial speech for voice privacy protection from personalized speech generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 411–11 415.

[18] J. Li, D. Ye, L. Tang, C. Chen, and S. Hu, "Voice guard: Protecting voice privacy with strong and imperceptible adversarial perturbation in the time domain." in *IJCAI*, 2023, pp. 4812–4820.

[19] C.-Y. Yang, S. G. Upadhyay, Y.-T. Wu, B.-H. Su, and C.-C. Lee, "Rw-voiceshield: Raw waveform-based adversarial attack on one-shot voice conversion," in *Proc. Interspeech 2024*, 2024, pp. 2730–2734.

[20] Z. Yu, S. Zhai, and N. 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.

[21] J. Gao, H. Li, Z. Zhang, and Z. Wu, "Black-box adversarial defense against voice conversion using latent space perturbation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[22] M. Zhang, B. Sisman, L. Zhao, and H. Li, "Deepconversion: Voice conversion with limited parallel training data," *Speech Communication*, vol. 122, pp. 31–43, 2020.

[23] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[24] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *International Conference on Learning Representations*.

[25] 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, 2022.

[26] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*.

[27] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang *et al.*, "Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 22 605–22 623.

[28] S. Dong, B. Chen, K. Ma, and G. Zhao, "Active defense against voice conversion through generative adversarial network," *IEEE Signal Processing Letters*, 2024.

[29] Q. Wang, J. Yao, Z. Sun, P. Guo, L. Xie, and J. H. Hansen, "Diff-fattack: Diffusion-based timbre-reserved adversarial attack in speaker identification," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[30] R. Ma, M. Qian, V. Raina, M. Gales, and K. Knill, "Universal acoustic adversarial attacks for flexible control of speech-llms," *arXiv preprint arXiv:2505.14286*, 2025.

[31] R. Li, Z. Liang, H. Zhang, T. Shi, Z. Cheng, J. Shi, C. Yang, and M. Tang, "Cloneshield: A framework for universal perturbation against zero-shot voice cloning," *arXiv preprint arXiv:2505.19119*, 2025.

[32] G. Chen, Y. Zhang, F. Song, T. Wang, X. Du, and Y. Liu, "A proactive and dual prevention mechanism against illegal song covers empowered by singing voice conversion," *arXiv preprint arXiv:2401.17133*, 2024.

[33] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.(2016)," *URL http://datashare. is. ed. ac. uk/handle/10283/2651*, 2016.

[34] M. Nussbaum, "Asymptotic equivalence of density estimation and gaussian white noise," *The Annals of Statistics*, pp. 2399–2430, 1996.

[35] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*.

[36] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber, "Natural evolution strategies," in *IEEE Congress on Evolutionary Computation (CEC 2008)*. IEEE, 2008, pp. 3381–3387.

[37] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, "Natural evolution strategies," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 949–980, 2014.