

F5E-TTS: ENHANCING SPEECH SYNTHESIS BY ALIGNING TEXT WITH RICH SEMANTIC REPRESENTATIONS

Yihang Chen¹ Hualei Wang¹ Na Li^{1*} Zhifeng Li*

¹ Tencent AI Lab, Shenzhen, China

cyanogen2003, hualei.wang, zhifeng0.li@gmail.com, lina011779@126.com

ABSTRACT

Modern Text-to-Speech (TTS) systems often produce intelligible but unnatural-sounding speech, a limitation we attribute to the semantic sparsity of text. To address this, we introduce **F5E-TTS**, a novel flow-matching framework that enriches the synthesis process by conditioning on both text and semantically-dense Phonetic PosteriorGrams (PPGs). The core of our framework is a training strategy where a Diffusion Transformer (DiT) backbone learns to implicitly align text and PPGs. This alignment is further encouraged by an explicit cross-modal regularization technique using a shared vector-quantized (VQ) codebook. This enriched content representation not only generates more natural-sounding speech but also significantly improves content accuracy. Our model achieves a state-of-the-art 16.4% relative Word Error Rate (WER) reduction over the strong F5-TTS baseline.

Index Terms— Text-to-speech, Semantic representation, Modality alignment

1. INTRODUCTION

While deep learning has propelled Text-to-Speech (TTS) to near-human intelligibility, synthesized speech often lacks the prosodic richness of human expression. We posit this ‘naturalness gap’ stems from a fundamental limitation: text is a semantically sparse representation of speech, devoid of the paralinguistic information required for a natural acoustic rendering. Our work aims to overcome this by enriching the synthesis process with a semantically dense modality, Phonetic PosteriorGrams (PPGs), which capture detailed phonetic and prosodic information from a reference utterance.

We introduce **F5E-TTS**, a non-autoregressive flow-matching framework [1] designed to leverage this richer signal. Our central hypothesis is that by training a model to generate speech from both sparse text and dense PPGs, it learns a more complete and robust internal representation of speech content. Our framework’s primary innovation is a training strategy where the model’s Diffusion Transformer (DiT) [2] backbone *implicitly* aligns text with PPGs.

To further enforce a shared understanding, we incorporate an explicit alignment using a shared vector-quantized (VQ) codebook [3], which encourages text and PPG representations into a common discrete latent space. The result is a model that generates more natural and expressive speech, with a significant corresponding improvement in content accuracy.

Our contributions are: 1) A novel TTS framework that improves naturalness by conditioning on both text and PPGs; 2) A training methodology that combines implicit alignment via a DiT backbone with an explicit VQ-codebook regularizer; 3) Demonstrating a new state-of-the-art 16.4% relative WER reduction over a strong F5-TTS [4] baseline.

2. RELATED WORK

Our work builds on non-autoregressive TTS models, known for efficient parallel generation. While much prior work has focused on improving intelligibility (WER) and speaker similarity (SIM-o), we focus on closing the semantic gap between text and its acoustic realization to enhance naturalness.

To bridge this gap, we draw inspiration from related speech processing fields. Phonetic PosteriorGrams (PPGs), often extracted using encoders like HuBERT [5], are a cornerstone of voice conversion (VC) for disentangling content from speaker identity [6]. We repurpose PPGs as a dense supervisory signal to teach a TTS model the prosodic nuances that text alone lacks. The concept of a shared vector-quantized (VQ) codebook was used effectively in the unified SpeechT5 [7] model to handle multiple speech and text tasks. Unlike such generalist models, we adapt the VQ codebook as a targeted alignment regularizer within a specialist TTS framework, specifically to unify sparse text and dense PPG representations. Finally, we employ Classifier-Free Guidance (CFG) [8] for user control, a standard technique for dynamically trading off synthesis objectives like speaker similarity and content accuracy at inference time. We also draw inspiration from recent work on controlling CFG strength for different conditions [9].

* Corresponding author.

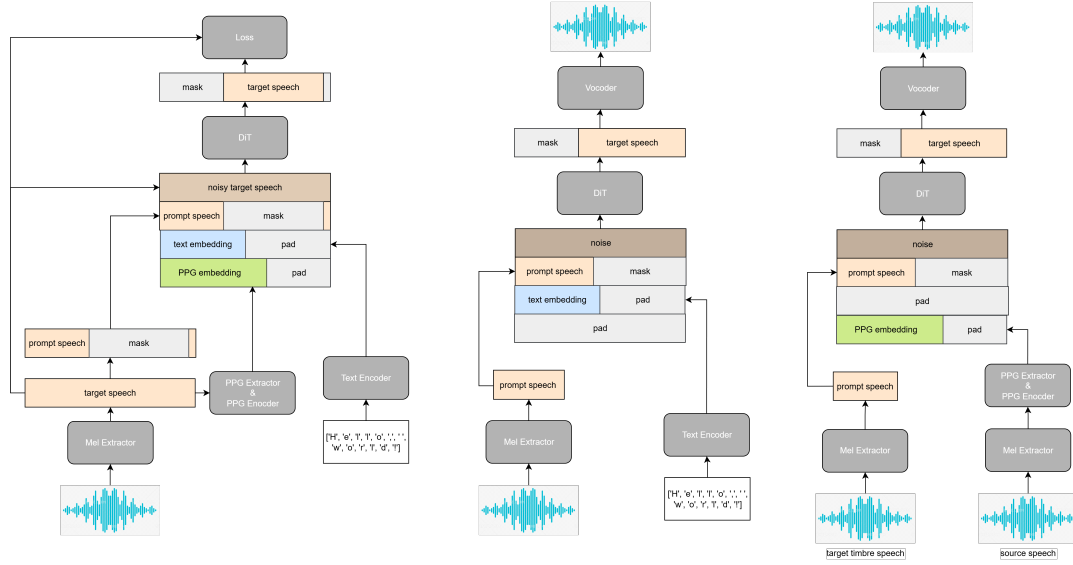


Fig. 1. The overall framework. During training (left), the model learns to align text and PPG modalities implicitly using DiT. At inference, it can perform zero-shot TTS (middle) or any-to-any voice conversion (right).

3. METHOD

We introduce a non-autoregressive framework designed to enhance TTS naturalness by enriching its semantic understanding of content.

3.1. Model Architecture

The core of F5E-TTS is a flow-matching network with a DiT [2] backbone. It takes four primary inputs: 1) a **text sequence**; 2) a **PPG sequence** for phonetic and prosodic detail; 3) a **prompt audio** mel-spectrogram for speaker timbre; and 4) a **noisy target** mel-spectrogram for the flow generation process [10]. Text and PPG sequences are processed by dedicated pre-nets, and their embeddings are padded to match the mel-spectrogram length. This combined sequence conditions the flow-matching module to generate the final mel-spectrogram, which is converted to a waveform using a pre-trained Vocos vocoder.

3.2. Training for Enhanced Content Representation

Our training strategy relies on a multi-task objective guided by Classifier-Free Guidance (CFG) [8], where conditional inputs are randomly dropped. This forces the model to learn robust and disentangled representations. The primary mechanism for alignment is the DiT backbone, which learns to map from both text and PPGs to the target speech.

3.2.1. Classifier-Free Guidance for Multi-Modal Learning

The primary training configurations are:

- **Primary TTS Task:** The model is conditioned on text and a speaker prompt, with PPGs dropped.
- **Rich Content Learning from PPGs:** The model is conditioned on PPGs and a speaker prompt, with text dropped. This VC-like objective forces the model to master the mapping from rich semantic features to speech, strengthening its ability to generate natural prosody.
- **Joint Alignment Task:** Both text and PPG inputs are provided, encouraging the DiT to implicitly learn a joint alignment between lexical and prosodic information.
- **Speaker Adaptation:** To further improve speaker modeling, we drop the prompt audio condition with a certain probability. This encourages the model to learn a generalized representation of speech that is less dependent on a specific speaker.

3.2.2. Explicit Alignment with a Shared Codebook

To explicitly align the textual and phonetic modalities into a unified semantic space, we introduce a cross-modal vector quantization method. This approach, inspired by SpeechT5 [7], builds a bridge between the two distinct data streams by mapping their continuous representations to a shared, discrete latent space.

Specifically, we employ a shared codebook, C^K , which contains K learnable embedding vectors. After the text and PPG inputs are processed by their respective pre-nets, the resulting continuous representations, u , are passed to a vector

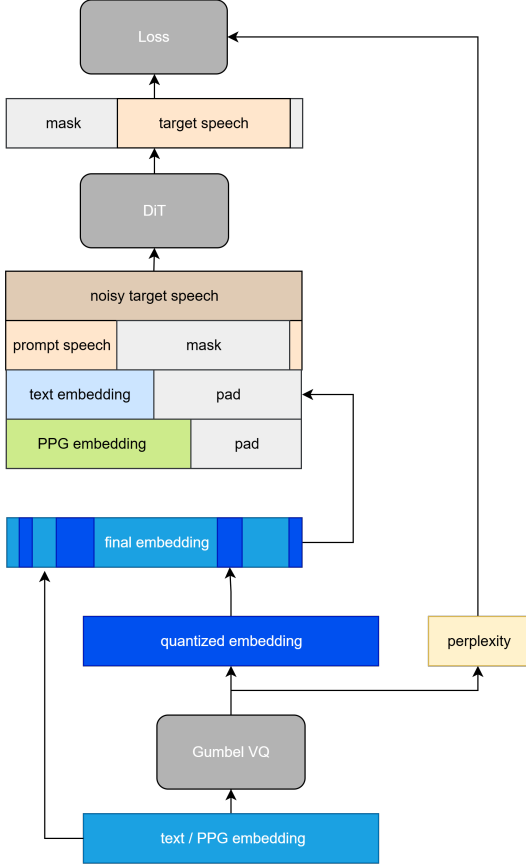


Fig. 2. The shared codebook aligns text and PPG representations into a unified semantic space.

quantizer. The quantizer converts each continuous vector u_i into a discrete codebook representation c_i by sampling using the Gumbel-Softmax [11]:

$$c_i = \sum_{j=1}^K \text{softmax}((\log \pi_j + g_j)/\tau) \cdot c_j \quad (1)$$

where π_j are the logits for each codebook entry, g_j are i.i.d. Gumbel samples, and τ is the temperature parameter. This operation is performed independently for both the text and PPG representations, but they both utilize the same codebook C^K , thereby forcing them into the same representational space.

To ensure the model learns a meaningful alignment and does not simply ignore the quantized representations, we randomly replace a certain percentage of the continuous vectors with their quantized counterparts before feeding them into the main DiT network.

To encourage the model to learn overlapping quantized vectors for both text and PPG tokens, thereby enhancing the alignment of different modalities into the same semantic space, we incorporate a perplexity loss, \mathcal{L}_p . This loss max-

imizes the entropy of the averaged softmax distribution of codebook usage across a batch:

$$\mathcal{L}_p = \frac{1}{K} \sum_{k=1}^K p_k \log p_k \quad (2)$$

where p_k is the averaged probability of choosing the k -th code in the codebook. The final pre-training objective is a weighted sum of the primary denoising loss and this perplexity loss:

$$\mathcal{L} = \mathcal{L}_{\text{denoising}} + \gamma \mathcal{L}_p \quad (3)$$

where γ is a hyperparameter balancing the two objectives, which we set to 0.1 during pre-training. Importantly, this cross-modal alignment mechanism is only active during training and does not introduce any additional computational overhead during inference, ensuring that the model’s performance at runtime remains efficient.

3.3. Controllable Inference

A key advantage of our framework is the ability to flexibly control the generation process at inference time. By leveraging the principles of CFG [8], we can independently adjust the guidance strength for each of the primary conditions: the speaker prompt (c_{spk}), the text (c_{txt}), and the PPGs (c_{ppg}).

The standard CFG formulation is extended to handle multiple conditions. The final guided output \hat{f}_θ is a weighted combination of the outputs from different conditional and unconditional outputs of the model. This method is inspired by recent work on controlling CFG strength for different conditions [9].

For the Text-to-Speech (TTS) task, the guided flow is:

$$\begin{aligned} \hat{f}_\theta(z_t, c_{\text{spk}}, c_{\text{txt}}, \emptyset) &= f_\theta(z_t, \emptyset, \emptyset, \emptyset) \\ &+ \alpha_{\text{txt}}[f_\theta(z_t, \emptyset, c_{\text{txt}}, \emptyset) - f_\theta(z_t, \emptyset, \emptyset, \emptyset)] \\ &+ \alpha_{\text{spk}}[f_\theta(z_t, c_{\text{spk}}, c_{\text{txt}}, \emptyset) - f_\theta(z_t, \emptyset, c_{\text{txt}}, \emptyset)] \end{aligned} \quad (4)$$

While not the focus of this work, our framework can theoretically perform voice conversion (VC) by conditioning on PPGs instead of text. The guided flow for this task would be:

$$\begin{aligned} \hat{f}_\theta(z_t, c_{\text{spk}}, \emptyset, c_{\text{ppg}}) &= f_\theta(z_t, \emptyset, \emptyset, \emptyset) \\ &+ \alpha_{\text{ppg}}[f_\theta(z_t, \emptyset, \emptyset, c_{\text{ppg}}) - f_\theta(z_t, \emptyset, \emptyset, \emptyset)] \\ &+ \alpha_{\text{spk}}[f_\theta(z_t, c_{\text{spk}}, \emptyset, c_{\text{ppg}}) - f_\theta(z_t, \emptyset, \emptyset, c_{\text{ppg}})] \end{aligned} \quad (5)$$

where α_{spk} , α_{txt} , and α_{ppg} are the respective guidance scales that can be adjusted by the user. This allows for a continuous trade-off between speaker similarity, faithfulness to the input text, and phonetic accuracy.

4. EXPERIMENTS AND RESULTS

4.1. Experiment Setup

We train a **Base model** (300M params) on the Emilia dataset [12] and a **Small model** (160M params) on the 585-hour LibriTTS dataset [13] for ablations. Audio is processed into 100-dim log mel-filterbanks at 24 kHz. We train on NVIDIA V100/A100/H20 GPUs with a batch size of 153,600 frames and use the AdamW optimizer with a $7.5e-5$ peak learning rate. We compare against a reproduced **F5-TTS** baseline [4]. We ablate our model by removing the shared codebook regularizer (**w/o CB**) and compare it to the full model (**w/ CB**). We evaluate zero-shot TTS performance using **WER** (Whisper-large-v3) [14], speaker similarity **SIM-o** (WavLM-large) [15], and naturalness via **UTMOS** [16].

4.2. Experiment Results

We trained Small models on LibriTTS to validate our design. As shown in Table 1, training with PPGs (w/o CB) improves WER over the baseline. Adding the shared codebook regularizer (w/ CB) yields further significant gains (e.g., WER from 2.29% to 1.91%). This confirms the benefit of our dual-modality alignment strategy, combining implicit learning with explicit regularization.

Table 1. Ablation study for Small models (160M) on LibriTTS. Our full model shows the best content accuracy (WER) while maintaining high speaker similarity (SIM-o).

Model	LibriSpeech-PC			seed-tts-eval EN		
	WER	SIM	UTMOS	WER	SIM	UTMOS
F5-TTS	2.20	0.600	TODO	2.29	0.577	TODO
w/o CB	2.11	0.603	TODO	2.08	0.581	TODO
w/ CB	2.05	0.607	TODO	1.91	0.588	TODO

Our Base model, trained on the Emilia dataset, significantly outperforms the strong F5-TTS baseline (Table 2). We achieve a relative WER reduction of **16.4%** (from 1.83% to 1.53%), demonstrating a marked improvement in intelligibility as a result of better semantic modeling.

Table 2. TTS results for Base models (300M) on seed-tts-eval EN. Our full model significantly improves WER over the baseline.

Model	WER(%) ↓	SIM-o ↑	UTMOS ↑
F5-TTS	1.83	0.670	TODO
w/o CB	1.65	0.652	TODO
w/ CB	1.53	0.637	TODO

Our model allows fine-grained control over the output. Tables 3 and 4 show that on our Base model, by adjusting

guidance scales (α_{spk} , α_{txt}), users can dynamically trade off between speaker similarity and content accuracy.

Table 3. Varying α_{spk} (speaker guidance) with fixed $\alpha_{\text{txt}} = 3.0$.

α_{spk}	WER(%) ↓	SIM-o ↑
2.5 (default)	1.49	0.626
3.0	1.55	0.639
3.5	1.66	0.640
4.0	1.74	0.637
4.5	1.77	0.630
5.0	1.94	0.621
5.5	2.31	0.608
6.0	2.58	0.593

Table 4. Varying α_{txt} (text guidance) with fixed $\alpha_{\text{spk}} = 3.5$.

α_{txt}	WER(%) ↓	SIM-o ↑
2.5	1.69	0.648
3.0 (default)	1.66	0.640
3.5	1.58	0.631
4.0	1.50	0.623
4.5	1.52	0.614
5.0	1.53	0.604
5.5	1.43	0.592
6.0	1.40	0.576

5. CONCLUSION

In this paper, we addressed the critical challenge of unnaturalness in modern TTS. We proposed **F5-TTS-E**, a novel framework whose main contribution is a training strategy that enriches content representation by conditioning a DiT-based flow model on both text and semantically dense PPGs. This approach relies on the model’s ability to *implicitly* align modalities, an effect which is reinforced with an *explicit* VQ-codebook regularizer. This focus on improving semantic understanding yielded dual benefits: more natural-sounding speech and vastly improved content accuracy, leading to a 16.4% relative WER reduction over a strong baseline.

While our primary goal was to enhance TTS, the model’s architecture inherently supports voice conversion, presenting a promising avenue for future research. This work demonstrates that focusing on the richness of semantic representation is a crucial step toward truly natural-sounding speech synthesis.

6. REFERENCES

- [1] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le, “Flow matching for generative modeling,” *International Conference on Learning Representations (ICLR)*, 2023.
- [2] William Peebles and Saining Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19033–19045.
- [3] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, “Neural discrete representation learning,” in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 6406–6415.
- [4] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, “F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching,” *arXiv preprint arXiv:2410.06885*, 2025.
- [5] Yu-An Chung, Wei-Ning Hsu, Haohan Wang, Chen-Yu Lee, Shinji Watanabe, and James Glass, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3466, 2021.
- [6] Cameron Churchwell, Max Morrison, and Bryan Pardo, “High-fidelity neural phonetic posteriorgrams,” *arXiv preprint arXiv:2402.17735*, 2024.
- [7] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al., “Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. 2022, Association for Computational Linguistics.
- [8] Jonathan Ho and Tim Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [9] Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al., “Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis,” *arXiv preprint arXiv:2502.18924*, 2025.
- [10] Jonathan Ho, Ajay N Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [11] Eric Jang, Shixiang Gu, and Ben Poole, “Categorical reparameterization with gumbel-softmax,” *International Conference on Learning Representations (ICLR)*, 2017.
- [12] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, and Yicheng Gu, “Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation,” *ResearchGate*, 2025.
- [13] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “Libri-tts: A corpus derived from librispeech for text-to-speech,” *arXiv preprint arXiv:1904.02994*, 2019.
- [14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023, vol. 202, pp. 28492–28518.
- [15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *arXiv preprint arXiv:2110.13900v5*, 2022.
- [16] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, “Utmos: Utokyo-sarulab system for voicemos challenge 2022,” in *Proceedings of Interspeech 2022*, 2022, pp. 4521–4525.