Frequency-Selective Boosting for CFM-based Speech Synthesis via Wavelet Decomposition

Anonymous EMNLP submission

Abstract

Conditional Flow Matching (CFM) models have advanced text-to-speech (TTS) synthesis, yet their efficiency and fidelity can be hampered by the uncoordinated evolution of spectral features during the generative ODE trajectory. Our analysis of DWT decomposition of the mel-spectrogram establishes that this incoordination between low-frequency (approximation) and high-frequency (detail) components often leads to unnecessary interference of subsequent iterations with the past developments and thus, demands prolonged iterations to achieve faithful speech. Furthermore, we demonstrate that directly adapting existing inference-time stabilization strategies, such as those inspired by MASF Qian et al., 2024 for diffusion models, exhibits poor generalizability to CFM-based TTS. This is due to fundamental differences in their generative dynamics, the time-varying reliability of intermediate clean data estimates in CFM, and potential mismatches with model-specific frequency evolution. To address these limitations, we propose a novel inference-time frequency-selective boosting strategy based on Wavelet decomposition, designed to explicitly enhance and synchronize the development of distinct mel-spectrogram frequency bands during the ODE solving process. Our experiments quantify significant improvements in the faithfulness and quality of generated audio, as measured by Fréchet Audio Distance (FAD), without any degradation in Word Error Rate (WER), showcasing a more robust and efficient path to high-quality speech synthesis in CFM models.

1 Introduction

Conditional Flow Matching (CFM) Lipman et al., 2023, particularly when guided by Optimal Transport (OT) principles, has significantly advanced text-to-speech (TTS) synthesis. This approach facilitates high-quality, parallel sampling through the integration of deterministic Ordinary Differential Equation (ODE) solvers. Prominent models, including Voicebox Le et al., 2023 and F5-TTS Chen et al., 2024, leverage this framework. They learn continuous generative trajectories where a neural network parameterizes a time-dependent vector field. This field defines an ODE that transforms samples from a simple prior distribution (e.g., Gaussian noise) to the complex speech data distribution. The target mel-spectrogram is synthesized by numerically solving this ODE; solvers approximate the continuous path by discretizing it into finite steps. In this CFM framework, OT principle is pivotal for establishing efficient target vector fields, often defining straight (linear) trajectories between a simple prior distribution and the target data distribution. The neural network then learns to parameterize this field by minimizing a Flow Matching objective. Such precise guidance along these simplified paths promotes stable training, leads to high-fidelity speech, and enables efficient inference. Voicebox leverages uniformly discretized ODE solvers to produce high-quality mel-spectrograms in a non-autoregressive manner. On the other hand, F5-TTS introduces a technique called Sway Sampling to accelerate inference. It employs a strategy that adaptively prioritizes certain time steps during ODE solving, focusing computational effort on stages most critical to the generation process.

Despite these innovations, our work identifies a key limitation in Section 3.1 shared across these CFM-based TTS models: an incoherent evolution of different frequency components within the mel-spectrogram x_t during the ODE integration. A similar problem was recognized by Yang et al., 2023 in the case of denoising diffusion models for image generation, where the low-frequency features of the image develop early on, while the high-frequency details start developing after a delay. Moreover, their work also suggests a limitation of diffusion denoising generative models to generate minor frequency components, which are

usually in the higher frequency ranges. In order to verify the existence of such an issue in CFM-based TTS models, we apply the Discrete Wavelet Transform (DWT) to analyze x_t and observe that its low-frequency (approximation) and high-frequency (detail) constituents often develop incoherently.

Wavelet Transforms (DWT/CWT) are established tools in audio and signal processing, with applications ranging from extracting speech parameters like pitch and formants Hamzenejadi et al., 2019 and enhancing mel-spectrograms for synthesis Hu et al., 2024, to creating robust features for tasks such as spoken language identification Dey et al., 2023 and analyzing biomedical signals like EEG data Goerttler et al., 2024. While these diverse applications underscore the versatility of wavelets in signal analysis and manipulation, our work introduces a distinct application. To the best of our knowledge, we are the first to employ DWT to specifically analyze and modulate the frequency sub-band evolution within the generative trajectory of Conditional Flow Matching (CFM) based text-to-speech models, addressing the internal dynamics of CFM generation in a novel way.

This spectral misalignment, persistent across various ODE solving strategies including adaptive sampling like Sway Sampling, often results in audible artifacts and generation inefficiencies. Such incoherent frequency development can destabilize the generative process, demanding more solver steps for perceptual convergence and thus hampering suitability for real-time deployment.

Previous work in image generation, such as the Moving Average Sampling in the Frequency Domain (MASF) technique Qian et al., 2024 for diffusion models, has demonstrated the utility of inference-time, frequency-specific smoothing to improve generative stability. Inspired by these principles, we investigated their applicability to CFM-based mel-spectrogram synthesis, including an analogous data-space projection to estimate clean mel-spectrograms from intermediate states. However, we found that directly adapting such inference-time smoothing strategies was insufficient for CFM models, explained in detail in Section 3.2.

To overcome these challenges, we introduce a novel frequency-selective boosting strategy in Section 3.3 that guides the training of CFM-based models by explicitly enhancing or suppressing the development of features in different sub-bands of the mel-spectrogram based on model-specific behavior. This targeted modulation fosters better synchronization of spectral components and encourages stable convergence throughout the generative process by boosting the contribution of lagging-behind components while penalizing any aggressive growth in other sub-bands.

We validate our approach across benchmarks using both Voicebox and F5-TTS frameworks, showing that our method improves perceptual audio quality as measured by Fréchet Audio Distance (FAD) Kilgour et al., 2019 while maintaining Word Error Rate (WER), ensuring that intelligibility is not compromised.

Key Contributions: The key contributions of this work are summarized below:

1. We establish that the CFM-based TTS models suffer (take longer iterations) to generate faithful speech because of the uncoordinated development of the approximation and detail features of the mel-spectrogram.

2. We demonstrate the poor generalizability of the MASF-based strategies for CFM-based models in stabilizing the generative process.

3. We propose a novel frequency-selective boosting strategy to enhance mel-spectrogram feature development in CFM-based TTS models.

4. We also quantify the improvement in faithfulness and quality of the generated audio using Fréchet Audio Distance (FAD), without any degradation in the Word Error Rate (WER).

2 Preliminaries

2.1 CNF/Flow Matching

Flow Matching with optimal transport continuous normalization flows (CNFs) Chen et al., 2018 provides a powerful framework for learning complex data distributions by transforming a simple prior distribution p_0 into a target data distribution p_1. This transformation is achieved through a time-dependent vector field $v_t : [0,1] \times R^d \rightarrow R^d$, which constructs a flow ϕ_t governed by the ODE:

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)); \phi_0(x) = x \qquad (1)$$

For a given flow $\phi_t : [0,1] \times R^d \rightarrow R^d$, we can derive the probability path $p_t(x)$ using the change

of variables formula:

$$p_t(x) = p_0(\phi_t^{-1}(x)) \left| \det \left(\frac{\partial \phi_t^{-1}}{\partial x}(x) \right) \right| \quad (2)$$

The vector field $v_t(x;\theta)$ parameterized by neural network θ can be trained with the Flow Matching objective:

$$\mathcal{L}_{FM}(\theta) = E_{t,p_t(x)} \left\| u_t(x) - v_t(x;\theta) \right\|^2 \quad (3)$$

where u_t is the vector field that generates p_t and $t \sim \mathcal{U}[0,1], \quad x \sim p_t(x)$. However, directly computing this objective is challenging, as we lack prior knowledge of p_t or v_t. Thus, a conditional probability path $p_t(x|x_1) = \mathcal{N}(x|x_1, \sigma_t^2 I)$, a Gaussian distribution centered at x_1 with a sufficiently small σ, is considered in actual training. The Conditional Flow Matching (CFM) objective is:

$$\begin{aligned} &\mathcal{L}_{CFM}(\theta) \\ &= E_{t,q(x_1),p_t(x|x_1)} \left\| u_t(x|x_1) - v_t(x;\theta) \right\|^2 \end{aligned} \quad (4)$$

The CFM loss is proved to have identical gradients with respect to θ. Here, x_1 is the random variable corresponding to training data. and μ_t and σ_t are the time-dependent mean and scalar standard deviation of the Gaussian distribution.

For leveraging the optimal transport (OT) path, which defines the conditional probability and vector field as:

$$p_t(x|x_1) = \mathcal{N}(x|tx_1, (1-(1-\sigma_{min})t)^2 I) \quad (5)$$

and

$$u_t(x|x_1) = \frac{x_1 - (1-\sigma_{min})x}{1-(1-\sigma_{min})t} \quad (6)$$

The OT path is particularly advantageous as it ensures points move with constant speed and direction, leading to more stable training and efficient inference. This choice simplifies the learning process while maintaining the model's expressive power.

2.2 Discrete Wavelet Transform

Wavelets are a class of special mathematical functions that are often used in the representation of data or other functions. Wavelets are often defined by a pair of functions consisting of a wavelet function and a scaling function, serving as a high-pass filter and a low-pass filter, respectively. Wavelet-based analyses and transforms process data at different *scales* or *resolutions* Graps, 1995.

Sub-Band	Frequency Axis	Time Axis
LL	Low-Pass	Low-Pass
LH	Low-Pass	High-Pass
HL	High-Pass	Low-Pass
HH	High-Pass	High-Pass

Table 1: Naming convention of the DWT sub-bands for the mel-spectrogram based on the direction of the low-pass and high-pass filters.

The 2D Discrete Wavelet Transform (2D-DWT) decomposes a two-dimensional signal $x[m,n]$ into four frequency sub-bands: x_{LL} (approximation), x_{LH}, x_{HL}, and x_{HH} (details), by applying separable low-pass and high-pass filtering along both dimensions followed by subsampling. This transform is invertible, allowing perfect reconstruction of the original signal from its sub-band coefficients via the inverse DWT (IDWT). The DWT and IDWT can be described as in (7).

$$\begin{aligned} \mathrm{DWT}(x) &= \{x_{LL}, x_{LH}, x_{HL}, x_{HH}\} \\ x &= \mathrm{IDWT}(x_{LL}, x_{LH}, x_{HL}, x_{HH}) \end{aligned} \quad (7)$$

Figure 1 illustrates the DWT decomposition of a mel-spectrogram (x). Table 1 shows the naming convention for the sub-bands used in this work.

3 Methodology

Our methodology first analyzes DWT decompositions of intermediate mel-spectrograms across the inference trajectory, revealing disordered frequency sub-band dynamics in CFM-based TTS models. We demonstrate that inconsistent growth rates between low-frequency (LL) and high-frequency (LH, HL, and HH) components degrade output quality, a problem mitigated in diffusion models but whose solutions fail to generalize to CFM architectures. To resolve this, we propose frequency-selective boosting, a lightweight compensation mechanism that explicitly coordinates sub-band evolution during generation while maintaining the CFM framework's efficiency.

3.1 Problem: Behavior of the Generative Process in CFM-based TTS Models

Figures 2 and 3 illustrate the development of ℓ^2-norm of the DWT sub-bands in mel-spectrograms during ODE function evaluations for F5-TTS and Voicebox respectively. In F5-TTS, the approximation (LL) sub-band develops rapidly from initialization, while high-frequency components (LH, HL,

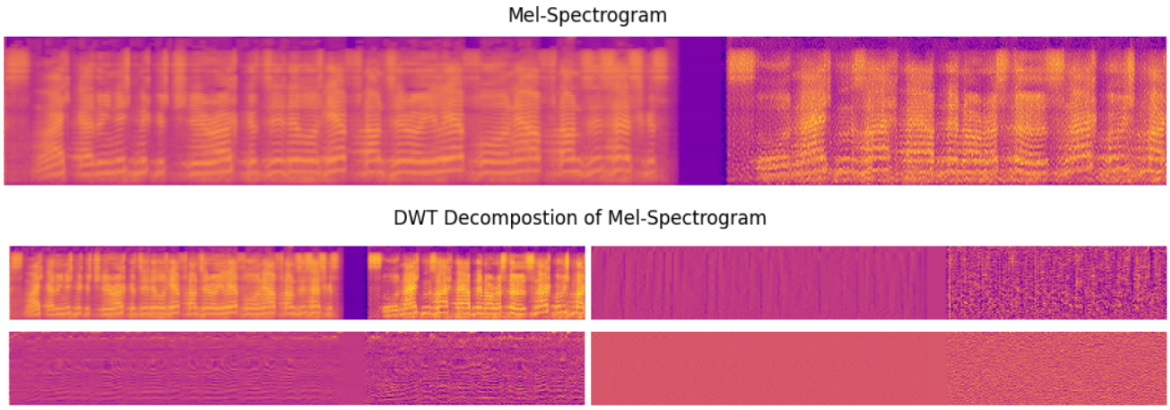

Figure 1: 2D-discrete wavelet transform of a mel-spectrogram. The original mel-spectrogram is shown on the top. The bottom four images show the DWT decomposition coefficients of the same mel-spectrogram. The approximation (LL) coefficient is shown at first, followed by the detail coefficients (LH, HL, HH).

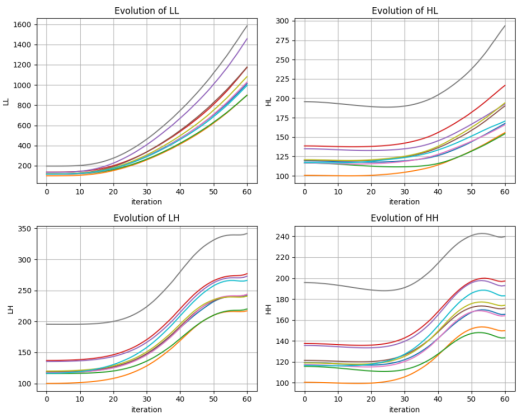

Figure 2: Evolution of the net content (ℓ^2-norm) of different frequency sub-bands obtained by discrete wavelet transform of the intermediate mel-spectrograms generated during various function evaluation steps (iteration) at the inference in F5-TTS. Each line corresponds to a different generative process.

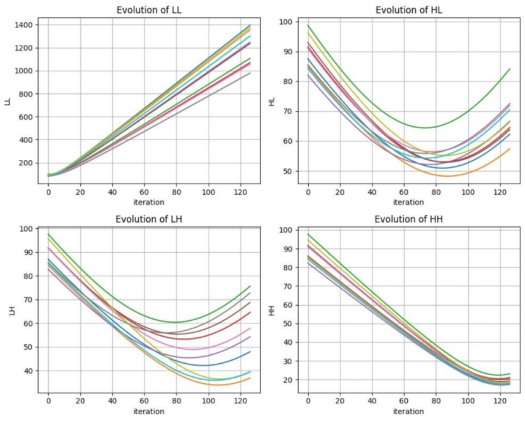

Figure 3: Evolution of the net content (ℓ^2-norm) of different frequency sub-bands obtained by discrete wavelet transform of the intermediate mel-spectrograms generated during various function evaluation steps (iteration) at the inference in Voicebox. Each line corresponds to a different generative process.

HH) emerge later. LH and HH sub-bands eventually saturate, while LL and HL continue steep growth. Contrastingly, Voicebox exhibits near-linear LL sub-band growth. LH and HL sub-bands show a fall and then delayed growth in the net content, while the contents of HH sub-band see a consistent fall.

Although models demonstrate consistent patterns in the respective sub-band developments across generative events, the uncoordinated evolution between approximation and detail features potentially increases computational complexity. This occurs when later developments in one sub-band disturb the contents of another sub-band, necessitating corrections to earlier developments, requiring additional function evaluations to achieve desired outputs. Consequently, this implies a poorer quality of generated speech than what could have been possible in the same number of iterations. Our experiments in Section 4 show that introducing a penalizing strategy to suppress aggressive changes while boosting the slower changes significantly improves the quality and faithfulness of the generated speech, supporting our analysis.

3.2 Stabilizing Mel-Spectrogram Generation with Frequency-Aware Trajectory Smoothing

fOur initial exploration into stabilizing mel-spectrogram generation in Conditional Flow Match-

ing (CFM) models drew inspiration from techniques successful in image diffusion, notably the Moving Average Sampling in Frequency domain (MASF) methodology Qian et al., 2024. MASF enhances the stability of Denoising Diffusion Implicit Models (DDIMs) Song et al., 2021 at inference time by first projecting noisy intermediate states to an estimate of the clean data, then decomposing this estimate into frequency sub-bands using Discrete Wavelet Transform (DWT), and subsequently applying frequency-specific Moving Averages combined with dynamic reweighting.

To adapt these principles to our CFM context, where mel-spectrograms x_t evolve along an Ordinary Differential Equation (ODE) trajectory ($t \in [0, 1]$, from noise x_0 to data x_1), the first step involved obtaining an analogous estimate of the clean target mel-spectrogram. For CFM models like F5-TTS and Voicebox, where the learned neural network $v_\theta(x_t, t)$ approximates the vector field $x_1 - x_0$, this data-space projection at time t is denoted by $\hat{x}_1^t$:

$$\hat{x}_1^t = v_\theta(x_t, t) + x_0 \tag{8}$$

where x_0 is the initial noise sample corresponding to the trajectory of x_t. The intention was then to explore the application of MASF-like inference-time smoothing mechanisms to the sequence of these $\hat{x}_1^t$ estimates derived at various points t along the ODE trajectory.

However, this approach of directly adapting MASF's inference-time smoothing strategies using the $\hat{x}_1^t$ sequence (from Eq. 8) did not yield the anticipated improvements in generation stability or quality for our CFM-based TTS models: it is noted that the characteristic of instability observed in Figure 2 is preserved in Figure 4. We identified several fundamental distinctions and challenges that render such a direct adaptation problematic:

Time-Varying Reliability of $\hat{x}_1^t$ Estimates: A key challenge when considering smoothing strategies for our Conditional Flow Matching (CFM) framework is the time-varying reliability of the data-space projected estimates, $\hat{x}_1^t = v_\theta(x_t, t) + x_0$. These $\hat{x}_1^t$ estimates are typically accurate representations of the target mel-spectrogram at early ODE stages ($t \approx 0$), as $v_\theta(x_t, t)$ is trained to map from initial noise towards clean data. However, $\hat{x}_1^t$ tends to degrade and distort as $t \to 1$ (when x_t approaches the target data x_1). This occurs because while the $\hat{x}_1^t$ formula relies on $v_\theta(x_t, t)$ predicting

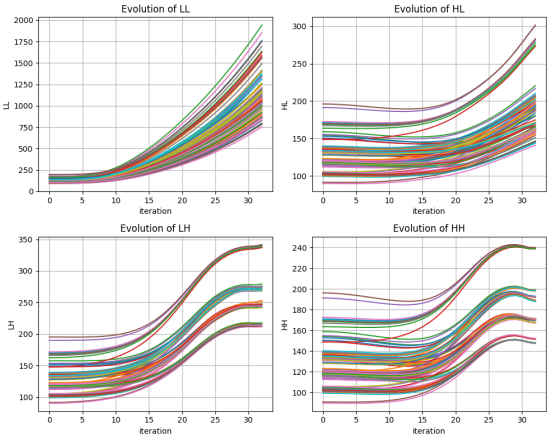

Figure 4: Dataspace moving average on F5TTS model does not yield the anticipated benefits

the total displacement $(x_1 - x_0)$, its role as the ODE velocity $\frac{dx_t}{dt}$ necessitates a diminishing magnitude for smooth convergence near x_1, causing the $\hat{x}_1^t$ estimate to become corrupted by adding x_0.

This behavior differs from estimates like DDIM's $\hat{x}_0^t$, where a noise prediction ϵ_θ within x_t, along with a precisely defined algebraic inversion based on a noise schedule $\bar{\alpha}_t$, naturally handles the progression towards the clean state without such systematic distortion of the estimate's target. Consequently, applying consistent smoothing (e.g., an Exponential Moving Average) across our sequence of $\hat{x}_1^t$ estimates, which vary significantly in reliability, is inherently problematic, as later, unreliable estimates can corrupt the smoothed average.

Mismatch with Model-Specific Mel-Spectrogram Frequency Dynamics: MASF utilizes pre-defined or linearly scheduled dynamic reweighting factors (e.g., $\beta_f(t)$ in its formulation) to modulate the influence of different frequency bands over its operational range. Such schedules, developed for image diffusion, may not align with the diverse, model-specific, and often non-linear evolutionary patterns of mel-spectrogram frequency components we observed in different CFM architectures. For example, low-frequency (LL) wavelet sub-bands exhibit quadratic-like energy growth in F5-TTS but a more linear trend in Voicebox (see Figure 3). A fixed or generic scheduling is unlikely to optimally cater to these distinct spectral dynamics in speech CFMs.

3.3 Framework: Model Specific Rescheduling

The sub-band norm evolution is plotted against the ODE solver steps. For every sub-band, we attempt

to fit a curve $f(t)$ and then define the reweighting coefficients $\beta_f(t)$ as $min(\delta_n, \frac{1}{\delta_d + f(t-t_0)})$. Motivation is to counteract the dynamics of $f(t)$ with coefficients of the form $1/f(t)$. The parameters δ_n, δ_d, t_0 are tuned by observing the sub-band norms with the aim of achieving saturation in the last few steps of the ODE. The clipping constant δ_n is necessary as repeated multiplication by a number greater than one(beyond a certain limit) will lead to divergence of the energy.

The frequency reweighting is performed in the DWT domain:

$$x^{t+1} = \text{IDWT}(\beta_f(t) \cdot \text{DWT}_f(x^t)); \quad \forall f \in \{\text{LL, HL, LH, HH}\} \tag{9}$$

While it is possible to fit multiple families of curves for the a given set of discrete points, not all families produce a stable output. Depending on the family, the reweighting may result in the energy either diverging or decaying to zero. Empirically, we evaluated polynomial, exponential, and logistic curve families. While polynomial curves showed high fitting accuracy, they often led to oscillatory or unstable rescaling, especially in later ODE steps. Exponential and clipped exponential forms were ultimately selected for their bounded growth and interpretability. The rescheduling curve is finally chosen through careful observation of the sub-band norms, such that it helps stabilise the evolution of the energies. The four sub-bands interact with and influence each other, but a degree of decoupling is assumed so that $\beta_f(t)$ can be tuned independently for each f. With this rescheduling, we aim to emphasize/de-emphasize the frequency content in the mel spectrogram at every iteration step, and target a synchronised convergence of the sub-band energies.

We observe that this reweighting may reduce the resulting norms of the sub-bands, especially the LL band. To ensure perceptual loudness is preserved post-rescheduling, we scale the final output mel-spectrogram by a global energy normalization constant derived from original dataset statistics.

This process can be performed for a multiple 'number of function evaluation' (NFE) values, and the rescheduling coefficient $\beta_f^N(t)$ will be tuned for different N. NFE is defined as the number of times the ODE is solved. Once $\beta_f^N(t)$ is parameterised over N, an explicit expression can be obtained through curve-fitting as will be demonstrated in Section 4.

4 Experiments and Analysis

We evaluate our proposed methodology on two distinct Conditional Flow Matching (CFM) based text-to-speech (TTS) models, operating on different languages to demonstrate broader applicability. For Hindi, we utilize an in-house implementation based on the Voicebox Le et al., 2023 architecture, trained on approximately 2k hours of Publically available Hindi speech data. This model comprises 103M parameters, featuring 12 layers and a 512-dimensional feed-forward hidden layer, and employs a standard deterministic ODE solver with fixed discretization steps for mel-spectrogram generation. We use the publicly available base F5-TTS model [1], which consists of approximately 350 parameters and is pre-trained on a multi-speaker English corpus with varied prosody as mentioned in (Chen et al., 2024). The F5-TTS model uniquely employs Sway Sampling to accelerate inference by adaptively prioritizing timesteps during ODE solving.

4.1 Experiments on F5-TTS

To assess real-world performance, we curated a test set of 100 utterances (10 diverse speakers/voice styles, 10 utterances each), recorded in various ambient environments on our campus. This dataset evaluates synthesis faithfulness for unseen speakers and prompts under these diverse conditions. As observed in Figure 2, the LL and HL bands tend to diverge towards the end of the process. LH and HH bands saturate and need not be rescheduled.

Qian et al., 2024 implements moving average in dataspace domain. We confirm with experiments on F5-TTS model that this technique does not help in improving results for CNF models and infact oberve a decline in performance (See figure 7). We therefore employ a rescheduling scheme in the DWT domain, without any projections onto the dataspace domain.

The LL band can be fit by an exponential curve and we choose the following template for $\beta_{LL}^N(t)$:

$$\beta_{LL}^N(t) = \exp\left(-\frac{t}{N(a_1 N + b_1)}\right) \tag{10}$$

for HL sub-band:

$$\beta_{HL}^N(t) = \exp\left(-\frac{t}{N(a_2 N^2 + b_2 N + c_2)}\right) \tag{11}$$

[1]Official github repo for code and checkpoint:https://github.com/SWivid/F5-TTS

We tune the rescheduling coefficients manually for a discrete set of N = {10, 20, 30, 45, 60} and by parameterising $\beta_f^N(t)$ as above, we can fit curves to obtain the values of (a_i, b_i, c_i).

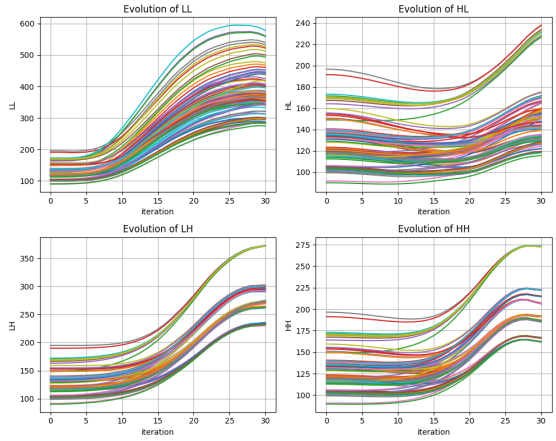

Figure 5: Sub-band norm evolution on F5-TTS model after frequency rescheduling

We observed that the sub-band norms (mainly LL) were saturating a few steps before the final step, and also explored the model performance by early-stopping the process. This 'knee point' beyond which the norms saturate, is empirically estimated:

$$\text{knee point} = \lfloor aN^2 + bN + c \rfloor \tag{12}$$

Thus, we are able to reduce the number of iteration steps that need to be computed, without any loss in performance.

4.2 Experiments on Voicebox

We employ a frequency scheduling scheme on Voicebox for NFE steps = 30. The coefficients, which are manually tuned, follow the following schedule(these are multiplied to x_f^t):

Sub-band	Reweighting schedule
LL	$(1 - (t/N) * 0.04)$
LH	$(1 - (t/N) * 0.0001)$
HL	$(1 - (t/N) * 0.03)$
HH	$\min(1.06, \frac{1}{(0.48+(t/N-30/32.0)^2)})$

Table 2: Reweighting schedule coefficients for Voicebox

For quantitative evaluation of the Voicebox model, particularly to measure Fréchet Audio Distance (FAD) over multiple utterances from single speakers across genders, we utilized test samples from the IndicTTS dataset Kumar et al., 2023. We selected samples from both male and female speakers for this analysis.

4.3 Evaluation Metrics

To evaluate the efficacy of our work, we quantitatively evaluate the generated audios on the following metrics:

1. Word Error Rate (WER): We utilized OpenAI's Whisper Radford et al., 2022 model to transcribe the generated audio files and calculate the WER between the transcribed text and the target text. WER is defined as

$$WER = \frac{S + D + I}{N} \tag{13}$$

where S is the number of substituted words in the transcribed text, D is the number of deleted words in the transcribed text, I is the number of inserted words in the transcribed text, and N is the total number of words in the target text. Consistently, we observed, audio generated by utilizing our strategy had no increase in the word error rate as compared to the speech generated by the original TTS model.

2. Fréchet Audio Distance (FAD): TTS models, apart from producing the correct words, also need to be faithful to the reference audio's speaking style. To verify the faithfulness of the speech, we employ FAD Kilgour et al., 2019, a metric which compares the difference in styles of groups of audio samples, giving a better quantization of styles Gui et al., 2024. Lower FAD scores correspond to similar audio styles. The audio files are converted into embeddings using the VGGish model Hershey et al., 2017, which converts the waveform into 128-D embedding representations of its semantic content. The embeddings of speeches under evaluation are put in one group, and those of the reference speeches are put in another group. A multivariate Gaussian distribution is fit over these groups, named $\mathcal{N}_e(\mu_e, \Sigma_e)$ and $\mathcal{N}_r(\mu_r, \Sigma_r)$ respectively. The FAD between these distributions is defined as:

$$\begin{aligned}\mathbf{F}(\mathcal{N}_e, \mathcal{N}_r) = &\|\mu_e - \mu_r\|^2 + \\ &\text{tr}(\Sigma_e + \Sigma_r - 2\sqrt{\Sigma_e \Sigma_r})\end{aligned} \tag{14}$$

4.4 Results

In this section, we present the results of the evaluation of two well-known CFM-based TTS models, F5-TTS Chen et al., 2024 and Voicebox Le et al., 2023, on the metrics described above.

4.4.1 F5-TTS

Simulation. The results are demonstrated for speech generation for NFE steps set to 32. We call

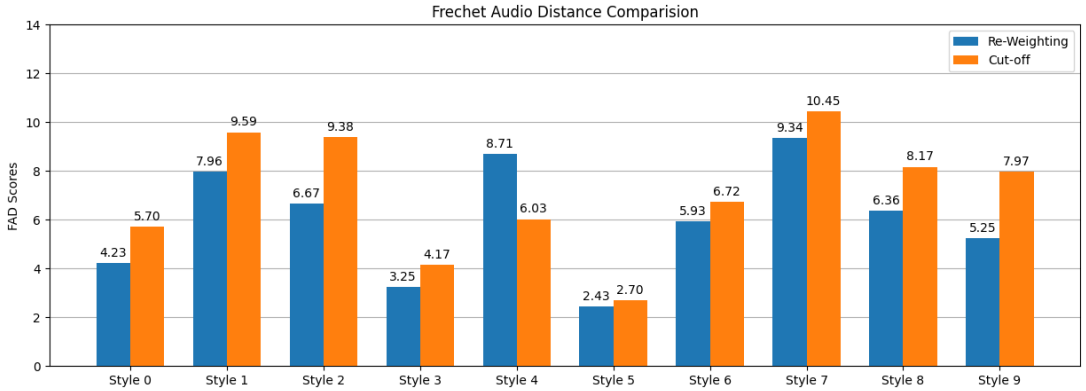

Figure 6: FAD score comparison between Frequency rescheduling early stop (Re-weighting Score) and baseline early stop (Cutoff Score) for F5-TTS model across 10 styles

the audio output of running the original model for 32 NFE steps the 'Original Full' audio. We observe the knee point at 25^{th} ODE evaluation. To evaluate, we construct the audio from the mel-spectrogram generated after 25^{th} NFE step in both the original and boosted model. The audio outputs of these are respectively named 'Original Cutoff' audio and 'Reweighted Cutoff' audio.

WER. We record no change in WER on our Reweighted Cutoff audio when compared to the Original Full audio.

FAD. We compare the following FAD scores:

1. Cutoff' score: FAD score between Original Cutoff audio and Original Full audio.

2. Re-weighting' score: FAD score between Reweighted Cutoff audio and Original Full audio.

These scores are calculated for 10 different styles of reference audios, namely Style 0–9, and 10 different speeches were generated for each style. The scores are compared in Figure 6. We can see that for 90% of the styles, we observe a significant decrease in the 'Re-weighting' score, when compared to the 'Cutoff' score, implying that the boosted model starts resembling the style of the original model faster.

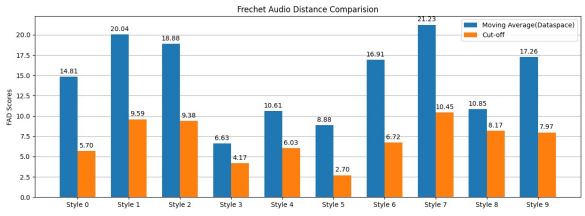

Figure 7: Comparison of moving average in data space (blue) based strategy with early cutoff (orange) of baseline model in F5-TTS. The significant increase in FAD scores signify that the method worsens the audio quality.

Input Style	Cutoff Score	Re-weighting Score
Female	5.85	**2.28**
Male	2.57	**1.79**

Table 3: FAD score comparison between Freqeucy rescheduling early stop (25/30, Re-weighting Score) and basline early stop (25/30, Cutoff Score) in Voicebox Hindi, across 2 styles

4.4.2 Voicebox

The models were run for NFE steps set to 30 to generate the audio. We extract the mel-spectrogram after the 25^{th} iteration, same as above, to compare the effects of boosting. The 'Cutoff' score is defined between the styles of generated audio from the early stopped baseline model and the original input samples. The 'Re-weighting' score is defined between the generated audio from the early-stopped boosted model and the original input samples. Table 3 depicts the respective scores computed using samples from the Hindi dataset Kumar et al., 2023.

5 Conclusion

Our method provides a principled inference-time strategy to improve sample efficiency in CFM-based TTS without any architectural modifications, enabling lightweight deployment. Apart from a reduction in solver steps that needed to be computed, no degradation in WER is observed. Additonally, we show a stronger similarity of our early-stopped model's output to the input audio styles(for Voicebox), compared to the early-stopped baseline model's output.

Limitations

Our current evaluation is limited to two languages—English (F5-TTS) and Hindi (Voicebox). While the proposed frequency-selective rescheduling generalizes across these models, further validation across morphologically rich and tonal languages remains pending. Additionally, the reweighting coefficients were tuned manually and may benefit from an automatic curve-fitting strategy. Finally, while inference time improved, we did not quantify real-time latency on constrained hardware. The strongest framework would be an adaptive strategy that would use the present-and-past-iteration sub-band norm values, their derivatives as well as automatic curve-fitting techniques.

References

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2018. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6572–6583, Red Hook, NY, USA. Curran Associates Inc.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching.

Spandan Dey, Premjeet Singh, and Goutam Saha. 2023. Wavelet scattering transform for improving generalization in low-resourced spoken language identification.

Stephan Goerttler, Fei He, and Min Wu. 2024. Balancing spectral, temporal and spatial information for eeg-based alzheimer's disease classification.

A. Graps. 1995. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2):50–61.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2024. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335.

Sajad Hamzenejadi, Seyed Amir Yousef Hosseini Goki, and Mahdieh Ghazvini. 2019. Extraction of speech pitch and formant frequencies using discrete wavelet transform. In *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, page 1–5. IEEE.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Guoqiang Hu, Huaning Tan, and Ruilai Li. 2024. A mel spectrogram enhancement paradigm based on cwt in speech synthesis.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A metric for evaluating music enhancement algorithms.

Gokul Karthik Kumar, Praveen S V, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. 2023. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. 2024. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8911–8920.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. 2023. Diffusion probabilistic model made slim. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22552–22562.