Harmonizing the Flow: Wavelet-Driven Frequency-Selective Boosting for CFM-Based TTS

Anonymous ACL submission

Abstract

Conditional Flow Matching (CFM) models for text-to-speech (TTS) suffer from incoherent frequency evolution during inference. While similar spectral imbalances are addressed in diffusion models for other domains, those generic solutions fail to generalize to the inherently uncoordinated acoustic dynamics of CFM. We demonstrate that this issue can be effectively mitigated by introducing a novel training-free frequency-selective boosting strategy. Using the Discrete Wavelet Transform (DWT), our method dynamically modulates mel-spectrogram sub-bands during ODE integration, synchronizing spectral development by penalizing aggressive low-frequency growth and boosting lagging high-frequency details. Validated across diverse architectures (Voicebox, Matcha-TTS, F5-TTS, IndicF5), our approach reduces the required Number of Function Evaluations (NFE) from 32 to 26 and improves Fréchet Audio Distance (FAD) by up to 61%, all without compromising the mean opinion score, speaker similarity and speech intelligibility.

1 Introduction

Conditional Flow Matching (CFM) (Lipman et al., 2023) is a powerful generative framework for text-to-speech (TTS) synthesis. It learns a continuous trajectory that transforms a simple prior distribution (e.g., a standard Gaussian) into a target data distribution over a normalized time interval. This transformation is governed by an Ordinary Differential Equation (ODE) whose underlying vector field is parameterized by a neural network (Chen et al., 2018). A key element of this framework is the use of Optimal Transport (OT) principles to define efficient, linear (straight-line) probability paths for this transformation. This direct guidance simplifies the vector field that the network must learn, promoting highly stable training and enabling high-fidelity speech generation.

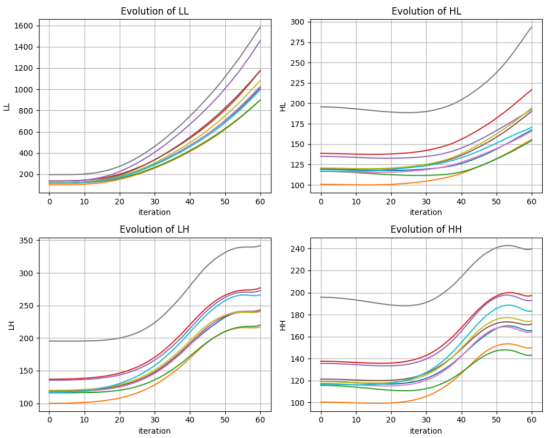

Figure 1: Evolution of DWT sub-band energy (ℓ^2-norm) across NFE steps during inference of F5-TTS.

To synthesize a mel-spectrogram during inference, the learned ODE is numerically solved by discretizing its continuous path into finite steps. While prominent models leverage this core Flow Matching mechanism, they differ in their solving strategies and computational trade-offs. For instance, Voicebox (Le et al., 2023) relies on standard **Uniform (Mid-point) ODE solvers** to achieve high-quality generative performance, typically requiring a larger number of evaluation steps. Capitalizing more aggressively on the efficiency of Optimal Transport, Matcha-TTS (Mehta et al., 2024) demonstrates that these straight-line trajectories can be resolved accurately using a simple, **Fixed-step Euler solver**, resulting in a lightweight architecture. In contrast to these uniform-step approaches, F5-TTS (Chen et al., 2024c) introduces an adaptive technique called **Sway Sampling**; this non-uniform strategy focuses computational effort on the most critical time steps in the generation process, accelerating inference by strategically distributing the solver's steps.

Despite recent innovations, we identify a critical, shared limitation in CFM-based TTS models

(see Section 4): the incoherent evolution of frequency components during ODE integration. As established by (Yang et al., 2023), continuous-time generative models are expected to exhibit a *frequency evolution* property, acting as a cascade of filters that operate in a "coarse-to-fine" manner to recover broad, low-frequency structures before transitioning to high-frequency details. However, neural networks also possess an intrinsic *frequency bias* (Rahaman et al., 2019; Yang et al., 2023). While heavily parameterized diffusion models may mitigate this imbalance through extensive reverse steps, (Yang et al., 2023) demonstrates that parameter-constrained networks fail to follow this spectral trajectory, leaving them unable to recover high-frequency components towards the end of generation. Our empirical analysis reveals that CFM-based acoustic models suffer from exactly this structural breakdown. Instead of a seamless spectral transition, frequency bias dominates; low-frequency components continue to grow throughout the entire integration, monopolizing the process and preventing the model from fully recovering fine acoustic details by the final step.

To investigate this phenomenon, following (Yang et al., 2023) and (Qian et al., 2024), we use the Discrete Wavelet Transform (DWT) (Graps, 1995) to analyze the l_2 norm of the time-evolution of the mel-spectrogram, x_t. Our analysis reveals an incoherent evolution between the low- and high-frequency components, a spectral misalignment that persists across diverse ODE solving strategies, from the uniform solvers in Voicebox to adaptive methods like Sway Sampling in F5-TTS (as shown in Figures 1 and 2, respectively). Specifically, we observe that the low-frequency coefficients continue to grow without saturation, providing an empirical manifestation of the aforementioned frequency bias.

While prior work in image generation, such as the MASF technique (Qian et al., 2024), uses inference-time, frequency-specific smoothing to address similar stability issues, we found that these principles do not readily transfer. As detailed in Section 4.2, a direct adaptation of such strategies proved insufficient to resolve the core instabilities in CFM-based models. To overcome these challenges, we introduce a novel **Frequency-Selective Boosting strategy** (Section 4.3). This method guides CFM by modulating mel DWT sub-bands based on model-specific behavior by boosting lagging components and penalizing aggressive growth

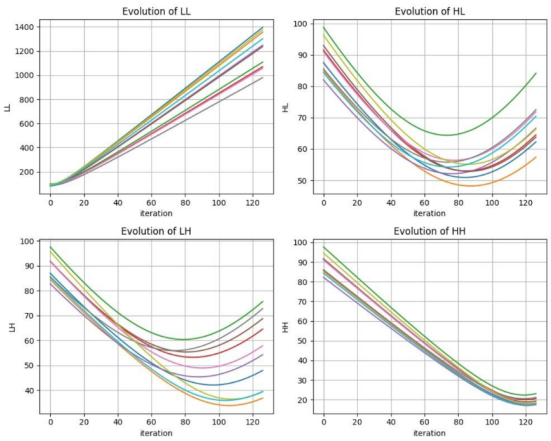

Figure 2: Evolution of DWT sub-band energy (ℓ^2-norm) across NFE steps during inference of Voicebox.

to synchronize spectral features and ensure stable convergence.

We validate our approach on F5-TTS, IndicF5, Voicebox, and Matcha benchmarks (Section 6), demonstrating improved perceptual audio quality (FAD) across diverse languages without compromising intelligibility (WER). Audio samples are available at the demo page[1]. To the best of our knowledge, we are the first to employ DWT to specifically analyze and modulate frequency sub-band evolution within the generative trajectory of CFM-based TTS, addressing the internal dynamics of mel-spectrogram generation in a novel way.

In summary, we have made the following contributions: 1) We establish that CFM-based TTS models suffer from uncoordinated development of the approximation and detail features of the mel-spectrogram and a severe frequency bias, making it difficult to recover high-frequency acoustic details. 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 feature development in CFM-based TTS models. We restrict the scope of this work to mel-spectrogram-based CFM-TTS models.

2 Preliminaries

2.1 CNF/Flow Matching

Flow Matching (FM) with continuous normalizing flows (CNFs) Chen et al., 2018 learns a vector field $v_t : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ that transforms a prior

[1]Anonymized Demo Page: https://anonymous.4open.science/w/boosed-cfm-demo/

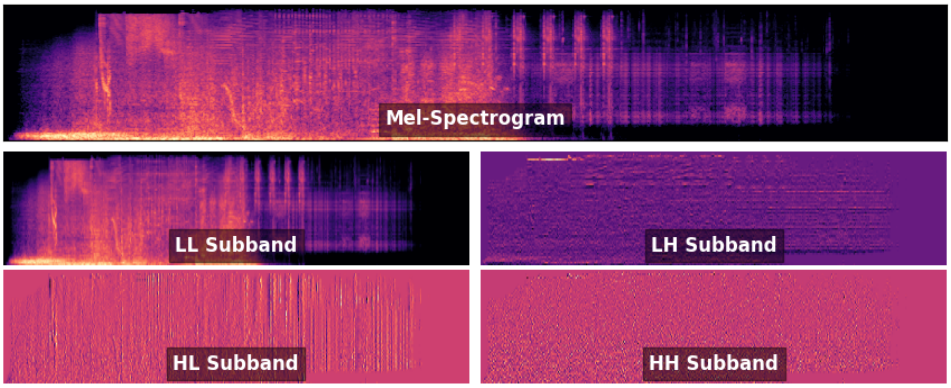

Figure 3: 2D-discrete wavelet transform of a mel-spectrogram. For the original mel-spectrogram shown at the top, the bottom four images show its Haar DWT subbands in the order LL (approximation), LH, HL, and HH (detail). In this figure, the LL subband is responsible for the overall structure, the HL subband serves as phoneme separation-based features, and the LH subband encodes changes in the frequency content of the speech.

distribution p_0 into a target distribution p_1 through the ODE:

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)), \quad \phi_0(x) = x \tag{1}$$

The FM objective is:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x\sim p_t}\|u_t(x) - v_t(x;\theta)\|^2 \tag{2}$$

In practice, Conditional Flow Matching (CFM) uses a Gaussian conditional path:

$$p_t(x|x_1) = \mathcal{N}(x|x_1, \sigma_t^2 I) \tag{3}$$

with objective:

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,q(x_1),p_t(x|x_1)}\|u_t(x|x_1) - v_t(x;\theta)\|^2 \tag{4}$$

Following Chen et al. (2018), we employ the optimal transport (OT) path:

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

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

The OT path improves training stability and inference efficiency.

2.2 Discrete Wavelet Transform

Wavelet analysis represents signals at multiple *scales* or *resolutions* using a pair of functions: a scaling function (low-pass filter) and a wavelet function (high-pass filter) Graps, 1995. The two-dimensional Discrete Wavelet Transform (2D-DWT) decomposes a signal $x[m,n]$ into four frequency sub-bands by applying separable low-pass and high-pass filtering along both dimensions, followed by subsampling. The resulting sub-bands are: x_{LL} (approximation) and x_{LH}, x_{HL}, x_{HH} (detail components). The transform is invertible, allowing perfect reconstruction of the original signal using the inverse DWT (IDWT).

In this work, we employ the Haar wavelet, whose 1D-filters are defined as $h_L = \frac{1}{\sqrt{2}}(1,1)$ and $h_H = \frac{1}{\sqrt{2}}(1,-1)$. The 2D-DWT sub-bands can equivalently be expressed as outer-products, $\mathbf{h}_{f_1 f_2} = \mathbf{h}_{f_1}\mathbf{h}_{f_2}^T$. Figure 3 illustrates the 2D Haar DWT decomposition of a mel-spectrogram (x).

3 Related Works

Wavelets and other multiscale transforms are established tools in audio and signal processing, with applications ranging from extracting speech parameters (Hamzenejadi et al., 2019) to enhancing mel-spectrograms for synthesis (Hu et al., 2024). In recent years, the multiscale representations have been used in neural speech synthesis for improving synthesis quality, reducing signal redundancy, and enabling faster training and inference (Al-Radhi et al., 2021; Zhang et al., 2024). Prior research demonstrates that learning wavelet-based features enhances the controllability and naturalness of synthesized speech (Al-Radhi et al., 2022; Wu and Hung, 2025), while similar frequency-aware fusion strategies have been explored in image generation (Chen et al., 2024a). Concurrently, multiscale "coarse-to-fine" architectures like FELLE (Wang et al., 2025), PixelFlow (Chen et al., 2025), and VALL-E (Chen et al., 2024b) have proven highly effective in boosting overall generative capabilities. However, while these foundational studies high-

light the utility of wavelets and multiscale representations, they primarily embed these concepts into the model architecture or training objective. In contrast, our work pioneers a fundamentally distinct application: repurposing wavelets as a lightweight, training-free intervention to dynamically steer and correct spectral evolution during inference.

4 Methodology

Our methodology employs DWT to expose the disordered frequency sub-band dynamics unique to CFM-based TTS models (see Section 4.1). After demonstrating that existing frequency-smoothing solutions from image diffusion fail to generalize to these acoustic models (Section 4.2), we introduce a novel frequency-selective boosting strategy: a lightweight, inference-time mechanism designed to explicitly coordinate and synchronize sub-band evolution during generation (Section 4.3).

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

In F5-TTS (see figure 1), the low-frequency approximation sub-band (LL) grows rapidly. While high-frequency detail components eventually emerge, some (LH, HH) saturate prematurely as the LL and HL bands continue a steep, unchecked ascent. Conversely, Voicebox and Matcha (see figure 2 and 7) exhibit near-linear LL growth alongside severely delayed development, or even a decaying magnitude, in their high-frequency sub-bands.

These observations reveal an *incoherent frequency evolution*: an uncoordinated development between spectral components coupled with a *frequency bias* towards dominant low-frequency content. Because the unchecked growth of the approximation sub-band severely delays the emergence of high-frequency details, the model struggles in the late stages of integration, making it exceedingly difficult to fully recover these lagging components. Prior research in spectral diffusion (Yang et al., 2023) demonstrates that coordinated frequency evolution is essential for stable generation, often prompting the use of explicit architectural priors to enforce this natural progression. Drawing on this foundational principle, we hypothesize that a dynamic, inference-time strategy is required to achieve high-fidelity speech generation with fewer steps in CFM frameworks. Specifically, explicitly restoring the coordination by penalizing low-frequency over-reliance and boosting lagging high-frequency components should resolve these instabilities. Our experiments in Section 6 validate this hypothesis, demonstrating that such targeted modulation significantly improves both speech quality and faithfulness.

4.2 The Limitations of Prior Frequency Smoothing

Our initial attempt to stabilize mel-spectrogram generation in CFM models was inspired by MASF (Qian et al., 2024), which stabilizes Denoising Diffusion Implicit Models (DDIMs) iteddim via frequency-domain smoothing. MASF estimates clean data from noisy states, decomposes with DWT, and applies frequency-specific moving averages with dynamic reweighting. For CFM-based TTS, we projected mel-spectrogram states x_t onto clean estimates $\hat{x}_1^t = x_t + (1-t)v_\theta(x_t, t)$. However, applying MASF-like smoothing to $\hat{x}_1^t$ sequences did not improve stability or quality, as instability persisted. We identified fundamental challenges to this direct adaptation:

Time-Varying Reliability of $\hat{x}_1^t$ Estimates: The $\hat{x}_1^t = x_t + (1-t)v_\theta(x_t, t)$ estimates behave differently across the ODE trajectory. As $t \to 1$, the weighting factor $(1-t)$ shrinks to zero, causing the estimator to rely almost entirely on the accumulated state x_t. If x_t contains integration errors from early spectral misalignments, the estimator inherits these artifacts directly. Unlike DDIM's noise prediction, where moving averages ensure stable convergence, uniform smoothing across $\hat{x}_1^t$ locks in late-stage errors and corrupts the generative trajectory.

Mismatch with Model-Specific Frequency Dynamics: MASF's generic schedules, designed for image diffusion, fail to stabilize CFM models due to their diverse and nonlinear frequency dynamics. For instance, the low-frequency growth is exponential in F5-TTS but linear in Voicebox, demonstrating why one-size-fits-all schedules are ineffective and why tailored, frequency-aware approaches are necessary.

4.3 Our Framework: Model Specific Reweighting

To address the incoherent frequency evolution described in Section 4.1, we propose a dynamic reweighting strategy that redistributes energy across the DWT sub-bands as the generation evolves. By appropriately scaling the sub-bands, we aim at improving emphasis on the other three sub-bands, and thus effectively reducing the em-

phasis on the LL sub-band towards the end. We propose that this gradual redistribution helps to achieve stable convergence in energy by the final NFE steps (see Figure 5).

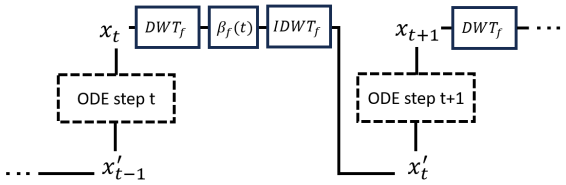

Figure 4: Reweighting process

To tailor the reweighting strategy, we define model-specific functions, $\beta_f(t)$ derived from a scaled and shifted form of the inverse of each sub-band's empirically observed energy curve, $1/c_f(t)$, allowing us to controllably attenuate dominant sub-bands while boosting energy-deficient ones.

$$\beta_f(t; c) = min\left(\delta_{clip}, \frac{1}{\delta_p + \delta_q c_f(t)}\right) \quad (7)$$

$$x'_t = \text{IDWT}(\beta_f(t) \cdot \text{DWT}_f(x_t)); \quad \forall f \in \{\text{LL, HL, LH, HH}\} \quad (8)$$

For a model, the curve of norm evolution $c_f(t)$ of different sub-bands is modeled from a family of curves including exponential, linear, and polynomial. δ_p, δ_q are the regularization terms and δ_{clip} is a clipping constant. Figure 4 demonstrates the reweighting process.

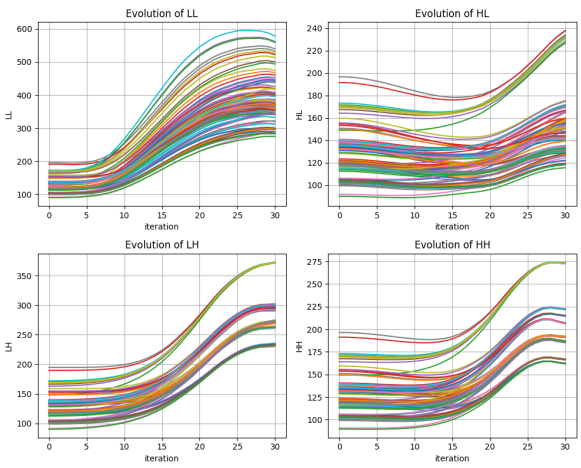

Figure 5: Sub-band norm evolution of F5-TTS model after frequency-specific reweighting

For **Reweighting F5-TTS**, we observe the trend of LL in Figure 1, which grows super-linearly with NFE step, the dominant LL sub-band can be attenuated using an exponential decay:

where $\alpha(N)$ is the decay term that controls the attenuation of LL growth. We generally realize it across the total number of function evaluations (N) by the following process. We choose a set of NFE steps, $\mathcal{S} = \{10, 20, 30, 45, 60\}$. For each $N \in \mathcal{S}$, we empirically tune the value of $\alpha(N)$. The coefficient $\beta^N_{LL}(t)$ falls exponentially with t and because $t \in \{1, .., N\}$ has different ranges for different Ns, we must normalize t by a function $q(N)$ that compensates for the different ranges. Hence, we find that a quadratic $q(N)$ is able to fit across all the values, resulting in the parametric form of $\alpha(N) = \alpha_0/q(N)$. We find that the energy of HL is sharply increasing even in the final iteration. To achieve stability by the final iterations, we reweight HL with the exponential form given in (9). Observing the trends in Figure 1, we note that LH and HH show a steady evolution with no large changes in their energies for the final 10 steps.

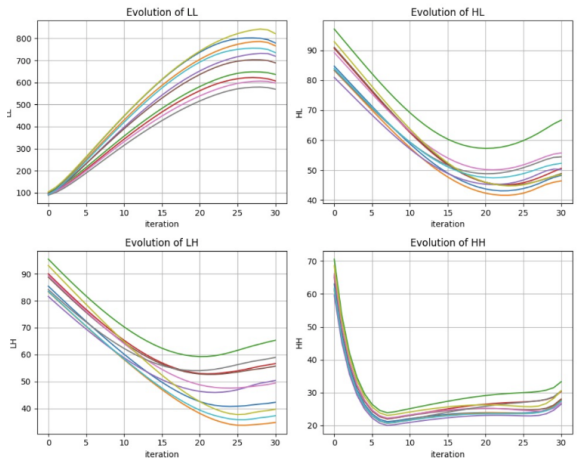

Figure 6: Sub-band norm evolution of Voicebox model after frequency-specific reweighting

$$\beta^N_{LL}(t) = \exp\left(-\alpha(N)t\right), \quad (9)$$

For **Reweighting Voicebox** we observe in Figure 2, LL shows a linear growth while the HH comp nent can be fit with a quadratic pattern. Thus, the reweighting parameters assume the form:

$$\beta_{LL}(t) = 1 - a * t/N \quad (10)$$

$$\beta_{HH}(t) = \min\left(\delta_{clip}, \frac{1}{s + \left(\frac{t-d}{\text{N}}\right)^2}\right) \quad (11)$$

As motivated in (7), we attempt to fit the HH evolution curve by a quadratic $c_{HH}(t)$ in this case.

Here, s and d are parameters of the quadratic equation which models the evolution and are chosen empirically, by means of a grid search.

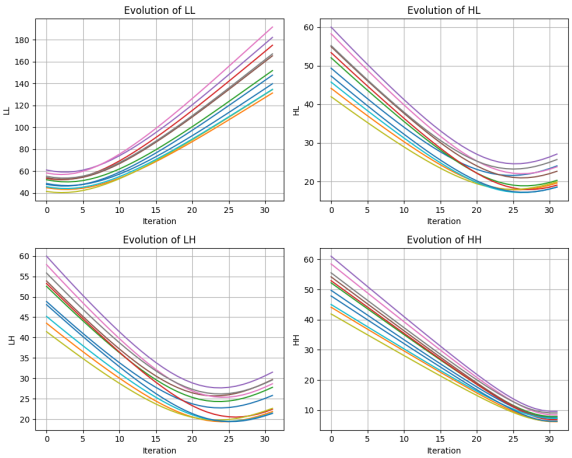

Figure 7: Sub-band norm evolution of MatchaTTS baseline model

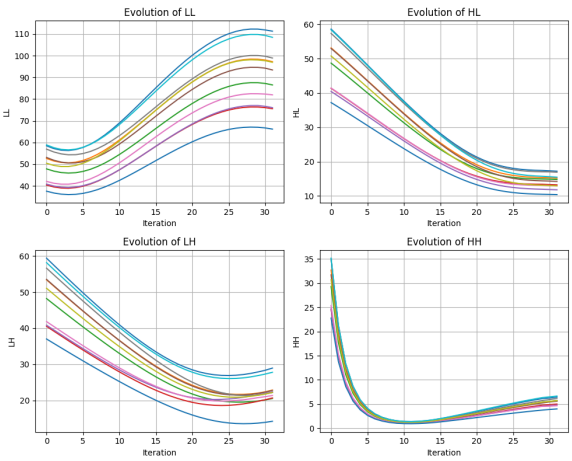

Figure 8: Sub-band norm evolution of MatchaTTS model after frequency-specific reweighting

For **Reweighting Matcha-TTS**, we observe in Figure 7 that the sub-band evolution profiles of the Matcha-TTS (Mehta et al., 2024) show similar trends as the Voicebox in Figure 2, thus the parametric form of reweighting coefficients for Voicebox, given in (10) and (11), are used for Matcha-TTS as well. We observe that reweighting the LH and HL sub-bands with a light slope,

$$\beta_{LH}(t) = (1 - b_{LH} * t/N) \tag{12}$$

$$\beta_{HL}(t) = (1 - b_{HL} * t/N) \tag{13}$$

For both these models, it is sufficient to achieve synchronicity. Here, b is a constant that is tuned empirically. The reweighted curves for Voicebox and Matcha-TTS are shown in Figure 6 and Figure 8, respectively.

5 Experimental Setup

5.1 Model Specifications

We pick the official checkpoint[2] for F5-TTS English, and IndicF5 (V et al., 2025) for Indic Languages, both consisting of approximately 350 parameters. For Voicebox, an in-house implementation based on the Voicebox (Le et al., 2023) architecture, comprising 150M parameters and trained on 3k hours of publicly available Hindi datasets. For Matcha-TTS, we also picked the official checkpoint[3], which is trained in the Single Female speaker LJ speech dataset (Ito and Johnson, 2017).

5.2 Evaluation Datasets

LibriTTS-clean and other (≈5k samples) each (Zen et al., 2019). Bengali and Tamil are drawn from IndicVoices-r (Sankar et al., 2024) 300 samples, expressing real-world scenarios. As Matcha-TTS is trained on a single female speaker, for speech synthesis, we only selected text from Librispeech Test set samples (≈400 samples) (for reproducibility purposes). For the Hindi test set that we used for Voicebox, we curated 100 real-life samples.

5.3 Evaluation Metrics

We quantitatively evaluate our method along three key dimensions: (1) *audio intelligibility*, (2) *perceptual quality*, and (3) *computational efficiency*.

Audio Intelligibility To measure how accurately the generated speech conveys the intended textual content, we compute the *Word Error Rate (WER)* and *Character Error Rate (CER)* using automatic speech recognition (ASR) system models. We specifically used Whisper Large v3 (Radford et al., 2022) for English and Indic Whisper (Bhogale et al., 2023) for Hindi, Tamil and Bengali.

Perceptual Quality We evaluate the acoustic perceptual quality of the generated speech using Fréchet Audio Distance (FAD) (Kilgour et al., 2019). To strengthen the evaluation, we evaluate the mean opinion score (MOS) and the speaker similarity score (SSS). FAD measures the distributional distance between multivariate Gaussians fitted to VGGish embeddings (Hershey et al., 2017) of the synthesized and reference audio datasets.

[2] https://github.com/SWivid/F5-TTS
[3] https://github.com/shivammehta25/Matcha-TTS

Model	Reweighting Strategy	Language	Dataset	NFE Steps	FAD ↓	WER ↓	CER ↓
F5-TTS	Baseline	English	LibriTTS-clean	32	0.99	0.066	0.037
	Boosted	English	LibriTTS-clean	26	**0.51**	0.067	0.038
	Baseline	English	LibriTTS-other	32	1.48	0.066	0.032
	Boosted	English	LibriTTS-other	26	**0.72**	0.066	0.033
Voicebox	Baseline	Hindi	Curated-Female	32	5.85	–	–
	Boosted	Hindi	Curated-Female	26†	**2.28**	–	–
	Baseline	Hindi	Curated-Male	32	2.57	–	–
	Boosted	Hindi	Curated-Male	26†	**1.79**	–	–
Matcha	Baseline	English	Librispeech	32	0.49	0.019	0.007
	Boosted	English	Librispeech	26	**0.45**	0.021	0.008

Table 1: Distributional Perceptual Quality Performance Comparison. Our frequency-selective boosting strategy dramatically improves perceptual fidelity while requiring 20% fewer inference steps (NFE) across diverse CFM-based architectures. (†Voicebox outputs evaluated via early stopping at step 26 of a 32-step schedule).

Computational Efficiency To assess practical deployment viability, we measure inference speed using average runtime and Real-Time Factor (RTF).

6 Results and Analysis

6.1 Performance Comparison: Enhancing distributional Quality with reduced NFE

Table 1 presents a comprehensive evaluation of our frequency-selective boosting strategy against standard baselines across multiple architectures (F5-TTS, Voicebox, and Matcha-TTS). To quantify perceptual quality and overall acoustic realism, we compute FAD between the generated outputs and the ground-truth audio.

Across all evaluated models and datasets, our boosting scheme consistently outperforms the baseline by achieving improvements in FAD scores while simultaneously reducing the NFE steps. Our boosted models reach perceptual convergence in just 26 NFE, compared to the 32 steps required by the baseline, resulting in approximately a 20% reduction in inference steps without any degradation in WER. This confirms that explicitly synchronizing sub-band evolution allows the ODE solver to find a high-fidelity trajectory much more efficiently.

6.2 Perceptual Quality: MOS and SSS

We compare the results of our boosted models with 26 NFE steps against the baseline models running till (1) 32 NFE steps and (2) stopped at 26 steps* for NFE = 32 in 2. The MOS for English audios was computed using UTMOS (Saeki et al., 2022), and the MOS scores for Tamil and Bengali audios were computed using IndicMOS (Udupa et al., 2024). The SSS between the generated audios and the original audios was computed using ECAPA-TDNN (Desplanques et al., 2020).

Matcha demonstrates that the boosted model at 26 NFE steps achieves a higher MOS (4.37) than both the 32-step baseline (4.31) and the 26-step baseline (3.63) for English speech. For F5-TTS (English, MOS), boosting at 26 steps clearly improves over the 26-step baseline and comes close to the 32-step baseline on LibriTTS Clean and Other (3.50 vs 3.64). In Indic-F5 (Tamil and Bengali, MOS), the boosted models consistently outperform the 26-step baseline and closely match the 32-step results. For SSS (English and Indic), boosted models slightly outperform or remain very close to both baselines, indicating competitive speaker similarity with fewer steps.

6.3 Ablation: Reweighting Schedules and Low-Resource Generalization

To validate our design choices for sub-band modulation, we perform an ablation study comparing exponential, linear, and degree-3 polynomial schedules for the approximation (LL) sub-band[4] (see Table 5). All outputs for this ablation were sampled at the 26th step out of the 32 NFE to rigorously test early convergence.

Our empirical results indicate that the **exponential schedule** yields the best performance, consistently driving the lowest FAD scores. The linear schedule performs reasonably well but falls slightly behind, suggesting our method is robust and not overly brittle to minor scheduling variations. However, the polynomial schedule fails to establish the necessary spectral synchronicity, resulting in se-

[4]Coefficients for other sub-bands (LH, HL, HH) are fixed across these experiments to isolate the effect of the LL reweighting.

Model	Dataset	Reweighting	NFE Steps*	MOS	SSS
Matcha	Librispeech English	Baseline	32	4.31	–
		Baseline	26	3.63	–
		Boosted	**26**	**4.37**	–
F5-TTS	LibriTTS Clean English	Baseline	**32**	**3.93**	0.76
		Baseline	26	3.05	0.74
		Boosted	26	3.76	**0.77**
F5-TTS	LibriTTS Other English	Baseline	**32**	**3.64**	0.72
		Baseline	26	2.69	0.71
		Boosted	26	3.50	**0.73**
Indic-F5	IndicVoices-r Tamil	Baseline	**32**	**2.35**	**0.83**
		Baseline	26	2.24	0.78
		Boosted	26	2.30	0.81
Indic-F5	IndicVoices-r Bengali	Baseline	**32**	**2.34**	**0.82**
		Baseline	26	2.07	0.75
		Boosted	26	2.26	0.78

Table 2: Best values within each dataset block are highlighted in bold. *Note:* Matcha uses a single speaker, so SSS is not applicable.

Language	WER	CER
Tamil	0.622	0.266
Bengali	0.652	0.541

Table 3: Ground Truth scores for IndicVoices-r.

Model Strategy	DWT Backend	RTF	Avg. Runtime
Baseline (32 NFE)	N/A	0.295	1.339s
Boosted (26 NFE)	`pywt` (CPU)	0.275	1.240s
Boosted (26 NFE)	`pytorch-wavelets` (GPU)	**0.248**	**1.119s**

Table 4: Inference efficiency evaluation on an NVIDIA H200 GPU.

Model	Reweighting Strategy	Language	Dataset	FAD↓	WER ↓	CER↓
F5-TTS	Baseline	English	LibriTTS-clean	12.93	**0.067**	**0.037**
	Exponential	English	LibriTTS-clean	**10.78**	0.072	0.047
	Linear	English	LibriTTS-clean	13.15	0.081	0.053
	Polynomial	English	LibriTTS-clean	17.29	0.198	0.133
	Baseline	English	LibriTTS-other	9.03	0.083	0.051
	Exponential	English	LibriTTS-other	**8.82**	**0.082**	**0.049**
	Linear	English	LibriTTS-other	10.78	0.083	0.050
	Polynomial	English	LibriTTS-other	14.36	0.431	0.304
Indic-F5	Baseline	Tamil	Indicvoices-r	7.58	**0.530**	**0.223**
	Exponential	Tamil	Indicvoices-r	**4.37**	0.531	0.230
	Baseline	Bengali	Indicvoices-r	9.30	0.511	0.396
	Exponential	Bengali	Indicvoices-r	**4.32**	**0.485**	0.457

Table 5: Ablation of LL sub-band reweighting schedules evaluated at 26 steps. The exponential schedule optimally counteracts natural low-frequency growth, yielding the highest perceptual quality across both high-resource (English) and low-resource (Indic) languages.

vere performance degradation. This confirms that the exponential penalty best counteracts the naturally aggressive low-frequency growth observed in CFM trajectories.

Generalization to Low-Resource Languages: We test our reweighting strategy on morphologically complex, low-resource Indic languages (Bengali and Tamil) using IndicF5 (V et al., 2025) (another model of the same Family as F5TTS). As shown in the bottom half of Table 5, our strategy achieves a remarkable improvement in generated audio quality (dropping FAD from 9.30 to 4.32 for Bengali). For contextual reference regarding the baseline intelligibility of these challenging datasets, Table 3 provides the Ground Truth WER/CER of the IndicVoices-r dataset.

6.4 Runtime Complexity and Real-World Viability

To assess computational overhead, we evaluated Real-Time Factor (RTF) and runtime on the IndicVoices-r Tamil test set using an NVIDIA H200 GPU (Table 4). By leveraging GPU-accelerated DWT (`pytorch-wavelets`), our 26-step boosted model achieves a **15% runtime reduction** over the 32-step baseline, confirming its efficiency for real-time deployment.

7 Conclusion

In this work, we identified and addressed the incoherent evolution of frequency components in CFM-based TTS models by introducing a novel, training-free frequency-selective boosting strategy. We demonstrated that our method generalizes to varying model architectures (VoiceBox, F5-TTS, Indic-F5, and Matcha) and languages (English, Hindi, Tamil, and Bengali), enhances generative stability, and achieves superior audio perceptual quality in fewer ODE steps with Mel-spectrogram based CFM-based TTS without any architectural modifications.

Limitations

Ultimately, our work establishes a lightweight empirical framework for controlling the internal generative dynamics of speech synthesis models. While the proposed frequency-selective rescheduling generalizes robustly across diverse architectures and datasets, the specific reweighting coefficients currently rely on empirical calibration tailored to model-specific behaviors. Developing an automated, adaptive curve-fitting strategy to dynamically infer these parameters represents a highly promising direction for future research. Lastly, this work validates the proposed training-free paradigm on mel-spectrogram-based CFM-TTS models. CFM models operating on other encoding domains can be studied as a future direction.

References

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Csaba Zainkó, and Géza Németh. 2022. Towards parametric speech synthesis using gaussian-markov model of spectral envelope and wavelet-based decomposition of F0. In *30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022*, pages 1150–1154. IEEE.

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Csaba Zainkó, and Géza Németh. 2021. Continuous wavelet vocoder-based decomposition of parametric speech waveform synthesis. In *Interspeech 2021*, pages 2212–2216.

Kaushal Santosh Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M Khapra, and Pratyush Kumar. 2023. Vistaar: Diverse benchmarks and training sets for indian language asr. *arXiv preprint arXiv:2305.15386*.

Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. 2024a. Frequency-aware feature fusion for dense image prediction. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):10763–10780.

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.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024b. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. 2025. Pixelflow: Pixel-space generative models with flow. *Preprint*, arXiv:2504.07963.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024c. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in time delay neural network. In *Proc. Interspeech 2020*.

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

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. *Preprint*, arXiv:2406.12164.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *Preprint*, arXiv:1812.08466.

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*.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345. IEEE.

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. *arXiv preprint*.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR.

Takaaki Saeki and 1 others. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Proc. Interspeech 2022*.

Ashwin Sankar, Srija Anand, Praveen Srinivasa Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *NeurIPS 2024 Datasets and Benchmarks*.

Sathvik Udupa and 1 others. 2024. Indicmos: Multilingual mos prediction for 7 indian languages. In *Proc. Interspeech 2024*.

Praveen S V, Srija Anand, Soma Siddhartha, and Mitesh M. Khapra. 2025. Indicf5: High-quality text-to-speech for indian languages.

Hui Wang, Shujie Liu, Lingwei Meng, Jinyu Li, Yifan Yang, Shiwan Zhao, Haiyang Sun, Yanqing Liu, Haoqin Sun, Jiaming Zhou, Yan Lu, and Yong Qin. 2025. Felle: Autoregressive speech synthesis with token-wise coarse-to-fine flow matching. *Preprint*, arXiv:2502.11128.

Zong-Tai Wu and Jeih-Weih Hung. 2025. Improving the speech enhancement model with discrete wavelet transform sub-band features in adaptive fullsubnet. *Electronics*, 14(7).

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.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proc. Interspeech*.

Xiangyu Zhang, Daijiao Liu, Hexin Liu, Qiquan Zhang, Hanyu Meng, Leibny Paola Garcia Perera, EngSiong Chng, and Lina Yao. 2024. Speaking in wavelet domain: A simple and efficient approach to speed up speech diffusion model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 159–171, Miami, Florida, USA. Association for Computational Linguistics.

8 APPENDIX

8.1 Hyperparamters

In this section, we list the values of all the hyperparameters defined in the paper.

F5-TTS, from (9), for $\beta^N_{LL}(t)$:

$$\alpha_0 = 0.7, q(N) = 0.3N^2.$$

For $\beta^N_{HH}(t)$:

$$\alpha_0 = -5.06,$$
$$q(N) = 0.02047N^3 + 0.8300N^2 - 2.15497N.$$

Voicebox, from (10) and (11), for $\beta^N_{LL}(t)$:

$$a = 0.04.$$

For $\beta^N_{LH}(t)$:

$$b_{LH} = 0.0001$$

For $\beta^N_{LH}(t)$:

$$b_{HL} = 0.03$$

For $\beta^N_{HH}(t)$:

$$\delta_{clip} = 1.06, s = 0.48, d = 30 \quad (\text{for } N = 32).$$

Matcha-TTS, from (10) and (11), for $\beta^N_{LL}(t)$:

$$a = 0.04.$$

For $\beta^N_{LH}(t)$:

$$b_{LH} = 0.0001$$

For $\beta^N_{LH}(t)$:

$$b_{HL} = 0.03$$

For $\beta^N_{HH}(t)$:

$$\delta_{clip} = 1.02, s = 0.8, d = 30 \quad (\text{for } N = 32).$$