GAF-Flow: An Auditory-Perceptual Flow Model for Enhanced Speech Synthesis

Rao Deng[1,2], Weike You[1], Linna Zhou[1,*], Hong Guo[2,*]
[1]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[2]Shanghai Key Laboratory of Forensic Medicine and Key Laboratory of Forensic Science, Ministry of Justice, Shanghai, China
{dengrao, ywk, zhoulinna}@bupt.edu.cn, guoh@ssfjd.cn

***Abstract*—This paper introduces GAF-Flow, a novel auditory-perceptual flow model for high-fidelity neural Text-To-Speech (TTS) synthesis. To address the limitations of "machine-centric" acoustic modeling, which often overlooks the non-linear frequency selectivity of human hearing, we propose three key innovations. First, a Gammatone-Mel Hybrid Filterbank is designed to mimic cochlear resolution and enrich high-frequency spectral details. Second, a Channel-wise Adaptive Normalization module harmonizes the heterogeneous statistical distributions between filter domains. Third, an auditory-prior supervision scheme provides targeted perceptual guidance during training. Experimental results on LJSpeech and VCTK datasets demonstrate that GAF-Flow significantly outperforms state-of-the-art baselines in both subjective (MOS, SMOS) and objective metrics (MCD, PESQ), successfully restoring fine spectral details and reducing artifacts while maintaining efficient inference. Our work bridges the gap between physiological auditory mechanisms and generative speech modeling, providing a perceptually-enhanced paradigm for high-quality synthesis.**

I. Introduction

The landscape of neural TTS has evolved rapidly, driven by the pursuit of high-fidelity and real-time synthesis. Early two-stage paradigms based on autoregressive modeling [10], [11], [23] achieved high quality but suffered from slow inference. To address this, non-autoregressive architectures [18], [19] and latent variable models like VITS [3] were introduced to enable parallel generation. Concurrently, GAN-based approaches [6], [8], [12], [24] and large-scale vocoders [31] utilized adversarial feedback to further sharpen spectral textures. Recently, generative modeling has witnessed a paradigm shift towards Denoising Diffusion Probabilistic Models (DDPMs) [2], [7], [26] and, more notably, Flow Matching frameworks [5], [25]. By simplifying generative trajectories, recent flow-based systems have achieved remarkable success in zero-shot voice cloning [27], [28] and prosodic modeling [22], [29], with scaling laws further pushing the boundaries of model capacity [9], [15], [17], [30].

However, a fundamental limitation persists across these state-of-the-art architectures: they largely operate on a "machine-centric" logic, assuming uniform statistical priority across all frequency bands. This design inherently overlooks the non-linear frequency selectivity of the human auditory system, particularly the transient cues critical for timbre perception, leading to a mismatch between signal-level optimization and perceptual quality. Auditory-guided metrics [1] and neuro-computational studies [21] affirm that aligning with auditory-nerve response patterns is a strong predictor of naturalness [9]. Yet, standard acoustic models fail to leverage these physiological priors. Consequently, the integration of biologically grounded auditory representations into generative speech models remains an open and crucial challenge.

Building on previous research, this paper proposes GAF-Flow, an auditory-perceptual flow model for speech synthesis. Our approach addresses the limitations of standard acoustic modeling through physiological auditory mechanisms and statistical domain alignment. The main contributions of this paper are summarized as follows:

- We propose a perceptual-quality-oriented synthesis framework based on the flow matching mechanism, which incorporates a Gammatone-Mel Hybrid Filterbank to mimic cochlear resolution and alleviate the information bottleneck in standard spectrograms.
- We introduce a Channel-wise Adaptive Normalization (CAN) module to harmonize the statistical distributions of heterogeneous feature domains, eliminating spectral fractures and ensuring coherent reconstruction.
- We design a dual-supervision scheme comprising Auditory Loss and Prior Regularizer, which provides targeted perceptual guidance to suppress artifacts and enhance naturalness.

II. Related work

A. Generative Flow Matching

Flow Matching (FM) has redefined the generative landscape by offering a simulation-free alternative to diffusion models, providing superior stability and convergence speed [32], [33]. In speech synthesis, Optimal Transport Conditional Flow Matching (OT-CFM) has been widely adopted to minimize the number of function evaluations while maintaining high fidelity [5], [34]. Recent works have scaled this paradigm to

Corresponding author: Linna Zhou and Hong Guo. This work was supported in part by the National Key Research and Development Program of China (Grant 2023YFC3305401), the National Natural Science Foundation of China (Grant 62172053 and Grant 62302059), and Shanghai Key Laboratory of Forensic Medicine and Key Laboratory of Forensic Science, Ministry of Justice (KF202420).

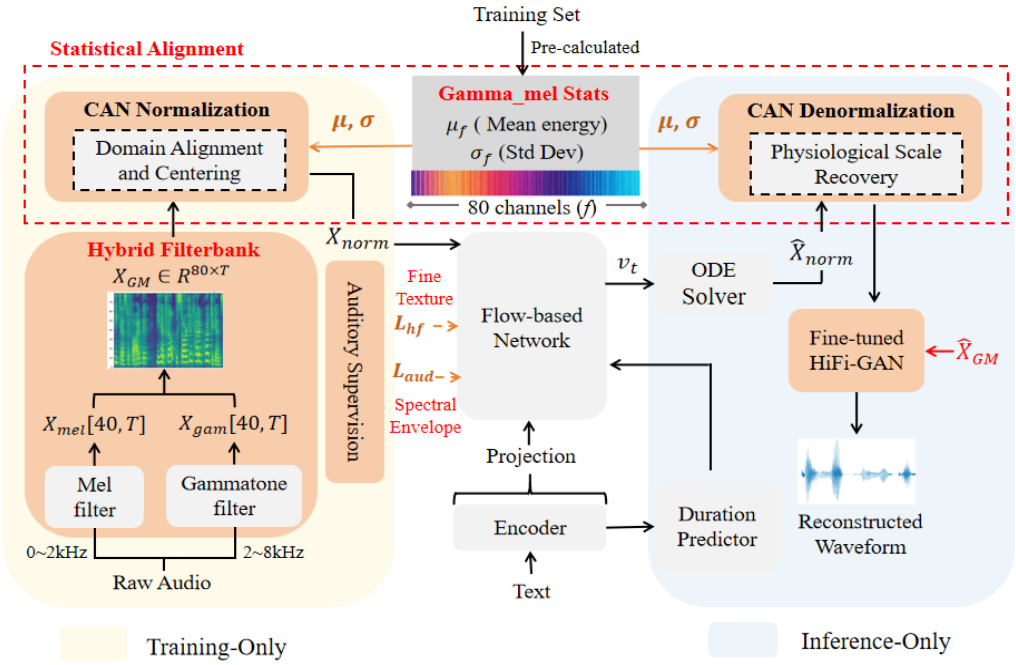


Fig. 1: Architecture of GAF-Flow, illustrating the training and inference stages bridged by shared core modules. Orange components highlight our primary contributions.

zero-shot scenarios, utilizing Rectified Flows [29] or Diffusion Transformers [27], [28] to capture complex speaker characteristics. These advancements underscore the effectiveness of flow-based models as a powerful backbone for modern, high-performance speech synthesis systems, establishing them as a leading approach for efficient and high-quality generation.

B. Auditory-Perceptual Modeling

Incorporating physiological priors into neural networks has proven effective for improving robustness in discriminative tasks. The Gammatone filterbank, which mimics the cochlea's non-linear frequency selectivity, has been extensively utilized in discriminative tasks such as robust Automatic Speech Recognition (ASR) and speech separation [13]. In the generative domain, recent neuro-computational frameworks like NeuroVoc [21] demonstrate the significant potential of auditory-inspired features [4], [12], showing that conditioning on simulated auditory nerve responses can enhance biological plausibility [14]. This growing body of work establishes auditory-based representations as a rich and valuable source of information, providing a strong foundation for their integration into more complex generative tasks like end-to-end TTS.

C. Feature Alignment and Normalization

Effective feature fusion is contingent upon robust normalization strategies. While standard techniques like Batch Normalization [38] and Layer Normalization [39] stabilize training via global statistics, the challenge of domain shifts has led to the development of adaptive methods. Techniques such as Adaptive Instance Normalization (AdaIN) [40] and Feature-wise Linear Modulation (FiLM) [41] have demonstrated remarkable success in dynamically modulating feature statistics based on external conditions, proving highly effective for style transfer and domain adaptation tasks. These adaptive normalization methods provide essential and powerful tools for managing heterogeneous data, showcasing their fundamental potential for enabling controlled and robust fusion within complex feature pipelines.

III. Method

As shown in Fig. 1, GAF-Flow is a perceptual-quality-oriented speech synthesis framework built upon Matcha-TTS [5]. We redesign its pipeline to operate in a physiologically grounded domain by introducing three key components: a bio-hybrid representation to restore cochlear resolution, a statistical alignment module to ensure stable generative paths, and perception-aware regularizers for targeted spectral supervision.

A. Gammatone-Mel Hybrid Filterbank

To enable auditory-aligned supervision, we first address the limitations of standard acoustic features. While standard Mel spectrograms effectively capture low-frequency tonal structures, their logarithmic compression severely degrades resolution above 2 kHz. Conversely, Gammatone filters, inspired by cochlear frequency decomposition [13], offer superior high-frequency resolution but are redundant at lower bands. To combine their complementary strengths, we propose a Frequency-split Hybrid Filterbank: Standard Mel filters for the low-frequency band ($<$2 kHz), Gammatone filters for the high-frequency band ($>$2 kHz).

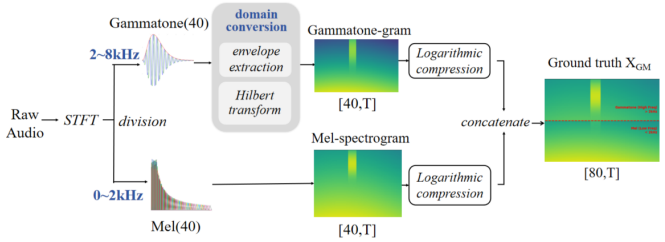


Fig. 2: System diagram depicting hybrid filterbank

As illustrated in Fig.2, the raw audio waveform is first transformed via Short-Time Fourier Transform (STFT). Then, the magnitude spectrum is split and processed through two parallel filterbank paths. Following filterbank analysis, both spectral streams undergo logarithmic compression $L = \log(|S|+10^{-5})$ before being concatenated along the frequency axis to form the unified hybrid spectrogram:

$$\mathbf{X}_{\text{GM}} = [\mathbf{X}_{\text{mel}}; \mathbf{X}_{\text{gamma}}] \in \mathbb{R}^{80\times T} \tag{1}$$

where T denotes the number of time frames. This 80-dimensional representation maintains exact compatibility with the Matcha-TTS interface, enabling seamless integration.

The unified hybrid spectrogram $\mathbf{X}_{\text{GM}}$ serves as the fundamental representation throughout the GAF-Flow pipeline. It acts as: (1) the ground-truth target for flow-matching-based acoustic modeling, (2) the primary domain for auditory-perceptual regularization, and (3) the training target for the fine-tuned HiFi-GAN vocoder. This End-to-End consistency ensures all components operate in a coherent, auditory-aligned domain, maximizing perceptual fidelity.

B. Channel-wise Adaptive Normalization

Mel and Gammatone features from the hybrid filterbank reside in heterogeneous spaces with divergent energy distributions, causing a statistical mismatch. Standard global normalization is dominated by high-energy Mel bands, suppressing fine-grained Gammatone textures and creating a statistical bias at their junction. This misalignment manifests as a spectral fracture—a sharp discontinuity at the 2kHz boundary (Fig. 3) that disrupts the continuity of the generative vector field. The fracture impedes coherent cross-domain learning due to conflicting gradient signals, leading to phase discontinuities and audible artifacts that degrade perceived naturalness.

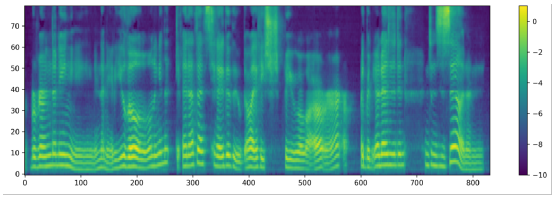


Fig. 3: Spectral fracture phenomenon

To establish a stable transport path, we propose Channel-wise Adaptive Normalization (CAN). CAN operates on a per-channel basis, computing independent mean μ_f and standard deviation σ_f for each frequency channel f across the training set. Given an input feature $y_{b,f,t}$ (where b, f, and t index the batch, frequency channel, and time frame, respectively), CAN applies the transformation:

$$\hat{y}_{b,f,t} = \frac{y_{b,f,t} - \mu_f}{\sigma_f + \epsilon}, \quad f \in \{1, \ldots, 80\} \tag{2}$$

where ϵ is a small constant for numerical stability.

This operation performs probability density alignment for each cochlear-scaled band, projecting both the Mel and Gammatone features onto a unified latent space with zero mean and unit variance per channel. By eliminating the inter-domain statistical bias, CAN ensures balanced contributions from all frequency bands to the vector field $v_\theta(x, t, c)$, effectively eliminating the spectral fracture and enabling the generation of seamless, high-fidelity spectrograms.

C. Auditory-Prior Supervision

Given the auditory-aligned representation, standard mathematical objectives like Mean Squared Error (MSE) remain perceptually undiscriminating. MSE treats all frequency bins with equal weight, failing to prioritize the specific bands where human hearing is most discerning or where generative artifacts are most audible. To this end, we introduce a dual-supervision scheme that prioritizes high-frequency fidelity.

1) Auditory Loss (L_{aud}): The Auditory Loss targets spectral precision in the high-frequency domain. It applies an L_1 constraint specifically to the Gammatone-derived bands $\mathcal{H}$ to enforce accurate reconstruction of fine spectral envelopes, prioritizing the acoustic textures essential for consonant clarity and natural timbre.

$$L_{\text{aud}}(y, \mu_y) = \frac{1}{|\mathcal{H}| \cdot \sum y_{\text{mask}}} \sum_{b,t} \sum_{f\in\mathcal{H}} |y_{b,f,t} - \mu_{y,b,f,t}| \cdot y_{\text{mask}} \tag{3}$$

where y is the ground-truth hybrid spectrogram and μ_y is the predicted encoder output. This targeted loss ensures the generator prioritizes fine-grained textures critical for consonant clarity and timbre.

2) Auditory Prior Regularizer (L_{hf}): The Auditory Prior Regularizer targets temporal coherence in the high-frequency domain. It imposes a smoothing constraint on the temporal evolution of the predicted features μ_y within bands $\mathcal{H}$, penalizing unnatural frame-to-frame fluctuations ("jitters") to ensure the generated spectro-temporal patterns are both sharp and stable.

$$L_{\text{hf}}(\mu_y) = \frac{1}{|\mathcal{H}| \cdot \sum y_{\text{mask}}} \sum_{b,t>0} \sum_{f\in\mathcal{H}} |\mu_{y,b,f,t} - \mu_{y,b,f,t-1}| \cdot y_{\text{mask}} \tag{4}$$

By penalizing unnatural temporal fluctuations in the Gammatone domain, this regularizer mimics the ear's sensitivity to temporal fine structures, leading to synthetic speech that achieves a higher degree of naturalness and clarity.

The final training objective integrates these auditory priors with the original flow-matching and duration losses:

$$L_{\text{total}} = L_{\text{dur}} + L_{\text{prior}}^{\text{orig}} + L_{\text{diff}} + \alpha L_{\text{aud}} + \beta L_{\text{hf}} \tag{5}$$

where α and β follow a linear warmup schedule: they remain 0 for the first 100 epochs, then linearly increase to their final values of 0.01 and 0.001, after which they remain constant. This schedule ensures training stability.

IV. Experiments

A. Experimental Setup

1) Datasets: We perform experiments primarily on the LJSpeech dataset [37], containing approximately 24 hours of audio clips from a single female speaker. The dataset is divided into training (12,500 samples), validation (100 samples), and testing (500 samples) sets, with the same split as [6], [8]. To evaluate cross-speaker generalization, we further utilize a subset of 20 speakers from the VCTK dataset [35]. For these speakers, 90% of their utterances are used for fine-tuning, and the remaining 10% for testing.

2) Preprocessing: All audio recordings are resampled to 22050 Hz, text inputs are phonemized using the *Phonemizer* library [16]. For feature extraction, spectrograms are computed using a Short-Time Fourier Transform (STFT) with an FFT size of 1024, a window length of 1024, and a hop size of 256. The frequency limits are set between $f_{min} = 0$ Hz and $f_{max} = 8000$ Hz.

3) Training: The model is trained using the AdamW optimizer ($\beta_1 = 0.8$, $\beta_2 = 0.99$) and a random seed of 1234 for reproducibility. The learning rate is initialized at 2×10^{-4} and follows an exponential decay of 0.999 per epoch, using a batch size of 16. To ensure high-fidelity waveform reconstruction from our specialized features, the HiFi-GAN vocoder [4] is fine-tuned specifically on the generated Gammatone-Mel hybrid spectrograms. All experiments are executed on a single NVIDIA RTX 4090 GPU.

B. Evaluation Metrics

In the subjective evaluation, we conduct a Mean Opinion Score (MOS) test with 30 listeners to evaluate naturalness (MOS-N) and speaker similarity (MOS-S). Each listener rated 20 randomly sampled utterances on a scale of 1 to 5. Results are reported as mean scores with 95% confidence intervals.

In the objective evaluation, we first conduct a Mel-cepstral distortion (MCD) [20] test to evaluate spectral distortion between synthesized and target speech. To directly assess the perceptual quality, we use the Perceptual Evaluation of Speech Quality (PESQ) to quantify listening quality due to its high correlation with MOS (r$\geq$0.9) [14]. Finally, we provide spectrogram comparisons for visual inspection of high-frequency fidelity.

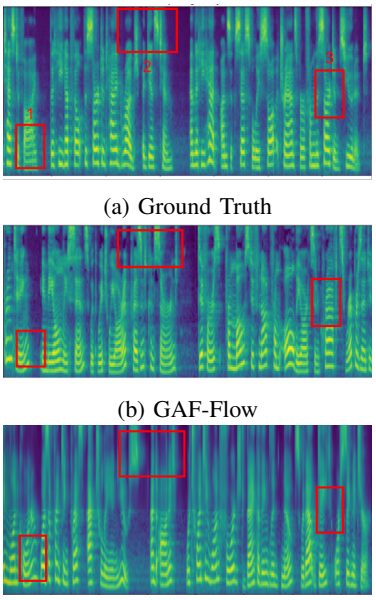
C. Results

1) Comparison with existing models:

The subjective evaluation results in Table I show that GAF-Flow achieves the best performance in both speech naturalness (MOS) and speaker similarity (SMOS) across datasets. It consistently surpasses the flow-based baseline Matcha-TTS, confirming the perceptual gains of our innovations. Notably, on the multi-speaker VCTK dataset, GAF-Flow maintains high SMOS, demonstrating robust generalization and its enhanced capability to preserve speaker identity. This is attributed to our bio-hybrid representation, which captures fine-grained physiological cues often lost in standard Mel-spectrograms.

The objective metrics provide complementary signal-level validation. GAF-Flow achieves the lowest MCD, indicating superior accuracy in spectral envelope reconstruction—a direct outcome of the CAN module's efficacy in aligning different statistical domains. Furthermore, its significant lead in PESQ scores validates the dual-supervision scheme. The proposed Auditory Loss (L_{aud}) and Prior Regularizer (L_{hf}) compel the generator to prioritize harmonic fidelity and temporal smoothness in critical high-frequency regions, effectively suppressing artifacts that degrade perceptual quality.

2) Visualizing Spectral Integrity:



(a) Ground Truth

(b) GAF-Flow

(c) Matcha-TTS

Fig. 4: Spectrogram comparison

To facilitate a more intuitive assessment of the synthesis quality, we visualize the spectrograms of test samples across different models in Fig. 4. Since baseline models employ varied Mel-scaling strategies that result in inconsistent visual scales, we focus on comparing GAF-Flow with the Ground Truth (GT) and the baseline Matcha-TTS.

A key observation is the visible spectral fracture in Matcha-TTS (Fig. 4c). This artifact arises from the statistical heterogeneity between Mel and Gammatone domains. In contrast, GAF-Flow (Fig. 4b) eliminates this discontinuity, demonstrating the CAN module's efficacy in feature alignment.

GAF-Flow shows superior acoustic detail preservation across the spectrum. In the low-frequency region (Box 1), GAF-Flow reconstructs sharper harmonic structures. This fidelity extends to the phonemic transition regions (Box 3), GAF-Flow replicates smoother energy shifts, avoiding the blurred transitions seen in the baseline. Most notably, in the high-frequency zone (Box 2), GAF-Flow maintains

TABLE I: Performance Comparison of Speech Synthesis on LJSpeech and VCTK Datasets. Bold values indicate the best performance among synthesized models.

Dataset	Method	MOS-N (↑)	SMOS (↑)	MCD [dB] (↓)	PESQ (↑)
LJSpeech	Ground Truth	4.41 ± 0.06	4.22 ± 0.12	-	4.12
	VITS	4.17 ± 0.12	3.92 ± 0.16	6.15	3.21
	Matcha-TTS	4.11 ± 0.09	4.03 ± 0.14	6.48	3.08
	Grad-TTS	4.06 ± 0.09	3.97 ± 0.14	6.92	2.88
	GAF-Flow (Ours)	$\mathbf{4.22 \pm 0.06}$	$\mathbf{4.09 \pm 0.09}$	**5.82**	**3.42**
VCTK	Ground Truth	4.38 ± 0.08	4.15 ± 0.11	-	4.01
	VITS	4.10 ± 0.11	3.82 ± 0.15	6.38	3.12
	Matcha-TTS	3.96 ± 0.10	3.84 ± 0.13	6.81	2.89
	Grad-TTS	3.88 ± 0.12	3.76 ± 0.14	7.24	2.72
	GAF-Flow (Ours)	$\mathbf{4.16 \pm 0.09}$	$\mathbf{3.98 \pm 0.10}$	**6.15**	**3.28**

fine-grained textures and transients that are critically over-smoothed in Matcha-TTS. These improvements stem from the synergistic effect of our bio-hybrid representation, which supplies richer spectral details, and the auditory-prior supervision, which ensures their accurate reconstruction during optimization.

3) Inference Efficiency:

We evaluate the inference efficiency of GAF-Flow by measuring the wall-clock inference time for generating utterances of varying lengths (from 30 to 1500 characters) on a single NVIDIA V100 GPU. The results are plotted on a logarithmic scale in Fig. 5, with trend lines indicating the scaling behavior ($O(L)$).

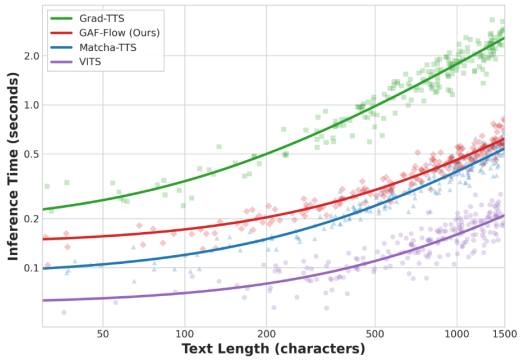


Fig. 5: Scatterplot of text length vs. inference time

As illustrated, the diffusion-based Grad-TTS exhibits the highest latency, while the fully parallel VITS model represents the lower bound for inference time. Crucially, GAF-Flow maintains inference speed nearly on par with Matcha-TTS, the flow-based baseline. The marginal additional latency is due to the computation of the hybrid features and the CAN module, but this overhead diminishes proportionally as utterance length increases. This demonstrates that our auditory-perceptual enhancements do not compromise the fast sampling advantage inherent to flow-matching models.

4) Ablation Studies:

To validate the individual necessity and synergistic effect of each proposed module, we conduct a comparative experiment on the LJSpeech dataset. The experimental results are shown in Table II. The most critical finding is the severe performance drop when removing the CAN module (MOS-N falls below 4.0), providing empirical evidence that the "spectral fracture" is highly detrimental to perceptual quality. The individual contributions of the auditory supervision terms are also clear: removing the Auditory Loss (L_{aud}) significantly reduces the PESQ score, indicating a loss of high-frequency spectral precision, while removing the Prior Regularizer (L_{hf}) leads to a slight degradation in MOS-N, corresponding to a loss of temporal smoothness. Finally, reverting to the Mel-only baseline leads to a comprehensive quality decrease, confirming the necessity of the hybrid representation.

TABLE II: Ablation Study Results

Configuration	MOS (↑)	PESQ (↑)
GAF-Flow	$\mathbf{4.22 \pm 0.06}$	**3.42**
w/o L_{aud} (Auditory Loss)	4.15 ± 0.08	3.26
w/o L_{hf} (Prior Regularizer)	4.19 ± 0.07	3.31
w/o CAN	3.98 ± 0.12	2.92
w.o GM representation (Baseline)	4.10 ± 0.10	3.10

V. Conclusion

In this paper, we propose GAF-Flow, an auditory-perceptual flow model that enhances naturalness and spectral fidelity in speech synthesis. Challenging the dominance of Mel-spectrogram-based acoustic modeling, we identify its limitation in capturing fine-grained high-frequency cues essential for timbre perception. To address this, we introduce a Gammatone-Mel Hybrid Filterbank to mimic cochlear resolution and propose Channel-wise Adaptive Normalization (CAN) to mitigate the spectral fracture caused by heterogeneous feature domains. Additionally, an Auditory-Prior Supervision scheme aligns the generative process with physiological hearing principles. Experimental results show that GAF-Flow outperforms state-of-the-art baselines, effectively eliminating metallic artifacts and restoring natural speech "airiness" without sacrificing inference efficiency. Future work will extend this bio-inspired framework to zero-shot voice cloning to enhance speaker similarity in unseen scenarios.

REFERENCES

[1] Ryosuke Sonobe, Takafumi Moriya, et al., "APG-MOS: Auditory perception guided-MOS predictor for synthetic speech," In ICASSP, pages 10461–10465. 2024.

[2] Zhifeng Kong, Wei Ping, et al., "DiffWave: A Versatile Diffusion Model for Audio Synthesis," In ICLR, 2021.

[3] Jaehyeon Kim, Jungil Kong, et al., "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," In ICML, pages 5530–5540. PMLR, 2021.

[4] Jungil Kong, Jaehyeon Kim, et al., "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," In NeurIPS, pages 17022–17033. 2020.

[5] Shivam Mehta, Ruibo Tu, et al., "Matcha-TTS: A fast TTS architecture with conditional flow matching," In ICASSP, pages 11341–11345. 2024.

[6] Yinghao Aaron Li, Cong Han, et al., "StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis," arXiv preprint arXiv:2205.15439. 2022.

[7] Vadim Popov, Ivan Vovk, et al., "Grad-TTS: A diffusion probabilistic model for text-to-speech," In ICML, pages 8599–8608. 2021.

[8] Yinghao Aaron Li, Cong Han, et al., "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," In NeurIPS. 2023.

[9] Zeqian Ju, Yuqiao Wang, et al., "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," arXiv preprint arXiv:2403.03100. 2024.

[10] Aaron van den Oord, Sander Dieleman, et al., "WaveNet: A generative model for raw audio," In SSW, page 125. 2016.

[11] Yuxuan Wang, R. J. Skerry-Ryan, et al., "Tacotron: Towards end-to-end speech synthesis," In Interspeech, pages 4006–4010. 2017.

[12] Ryuichi Yamamoto, Eunwoo Song, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," In ICASSP, pages 6199–6203. 2020.

[13] M. Dörfler, T. Grill, et al., "Basic filters for convolutional neural networks applied to music: Training or design?" Neural Computing and Applications, vol. 32, pp. 941–954, 2020.

[14] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," In IEEE TASLP, vol. 16, no. 1, pages 229–238. 2008.

[15] Chengyi Wang, Sanyuan Chen, et al., "Neural codec language models are zero-shot text to speech synthesizers," arXiv preprint arXiv:2301.02111. 2023.

[16] Mathieu Bernard and Hadrien Titeux, "Phonemizer: Text to phones transcription for multiple languages in Python," In JOSS, vol. 6, no. 68, page 3958. 2021.

[17] Kai Shen, Zeqian Ju, et al., "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," arXiv preprint arXiv:2304.09116. 2023.

[18] Yi Ren, Yangjun Ruan, et al., "FastSpeech: Fast, robust and controllable text to speech," In NeurIPS. 2019.

[19] Yi Ren, Chenxu Hu, et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," In ICLR. 2021.

[20] James C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," In IEEE Transactions on Automatic Control, vol. 37, no. 3, pages 332–341. 1992.

[21] Jacob de Nobel, Jeroen J. Briaire, et al., "From spikes to speech: NeuroVoc – A biologically plausible vocoder framework for auditory perception and cochlear implant simulation," In Hearing Research, vol. 469, page 109469. 2025.

[22] Haoyu Wang, Sizhe Shan, et al., "ProsodyFlow: High-fidelity text-to-speech through conditional flow modeling and prosody modeling with large speech language models," In Proceedings of COLING, pages 7748–7753. 2025.

[23] Jonathan Shen, Ruoming Pang, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," In ICASSP, pages 4779–4783. 2018.

[24] Won Jang, Dan Lim, et al., "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," In Interspeech, pages 2207–2211. 2021.

[25] Zhen Ye, Wei Xue, et al., "CoMoSpeech: One-step speech and singing synthesis via consistency distillation," In ACM Multimedia, pages 1831–1839. 2023.

[26] Rongjie Huang, Zhou Zhao, et al., "ProDiff: Progressive fast diffusion model for high-quality text-to-speech," In ACM Multimedia, pages 2595–2605. 2022.

[27] Yushen Chen, Zhikang Niu, et al., "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," arXiv preprint arXiv:2410.06885. 2024.

[28] Sefik Emre Eskimez, Xiaofei Wang, et al., "E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS," arXiv preprint arXiv:2406.18009. 2024.

[29] Hyun Joon Park, Jeongmin Liu, et al., "RapFlow-TTS: Rapid and High-Fidelity Text-to-Speech with Improved Consistency Flow Matching," In Interspeech. 2025.

[30] Xu Tan, Jiawei Chen, et al., "NaturalSpeech: End-to-end text to speech synthesis with human-level quality," In NeurIPS. 2022.

[31] Sang-gil Lee, Wei Ping, et al., "BigVGAN: A universal neural vocoder with large-scale training," In ICLR. 2023.

[32] Yaron Lipman, Ricky T. Q. Chen, et al., "Flow Matching for Generative Modeling," In Machine Learning. 2023.

[33] Xingchao Liu, Chengyue Gong, et al., "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow," In Machine Learning. 2022.

[34] Yiwei Guo, Chenpeng Du, et al., "VoiceFlow: Efficient Text-to-Speech with Rectified Flow Matching," In ICASSP. 2024.

[35] Christophe Veaux, Junichi Yamagishi, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[36] Pranay Manocha, Adam Finkelstein, et al., "A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences," In ISMIR. 2021.

[37] Keith Ito and Linda Johnson, "The LJ Speech Dataset," 2017.

[38] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," In ICML. 2015.

[39] Jimmy Lei Ba, Jamie Ryan Kiros, et al., "Layer Normalization," arXiv preprint arXiv:1607.06450. 2016.

[40] Xun Huang and Serge Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," In ICCV. 2017.

[41] Ethan Perez, Florian Strub, et al., "FiLM: Visual Reasoning with a General Conditioning Layer," In AAAI. 2018.