FNH-TTS: A Fast, Natural, and Human-Like Speech Synthesis System with advanced prosodic modeling based on Mixture of Experts

1st Qingliang Meng
Megatronix
Beijing, China
qingliang.meng@megatronix.co

2nd Yuqing Deng
Megatronix
Beijing, China
yuqing.deng@megatronix.co

3rd Wei Liang
Megatronix
Beijing, China
wei.liang@megatronix.co

4th Limei Yu
Megatronix
Beijing, China
limei.yu@megatronix.co

5th Huizhi Liang
Newcastle University
Newcastle upon Tyne, United Kingdom
huizhi.liang@newcastle.ac.uk

6th Tian Li*
Newcastle University
Newcastle upon Tyne, United Kingdom
huizhi.liang@newcastle.ac.uk

***Abstract*—Achieving natural and human-like speech synthesis with low inference costs remains a major challenge in speech synthesis research. This study focuses on human prosodic patterns and synthesized spectrum harmony, addressing the challenges of prosody modeling and artifact issues in non-autoregressive models. To enhance prosody modeling and synthesis quality, we introduce a new Duration Predictor based on the Mixture of Experts alongside a new Vocoder with two advanced multi-scale discriminators. We integrated the these new modules into the VITS system, forming our FNH-TTS system. Our experiments on LJSpeech, VCTK, and LibriTTS demonstrate the system's superiority in synthesis quality, phoneme duration prediction, Vocoder results, and synthesis speed. Our prosody visualization results show that FNH-TTS produces duration predictions that more closely align with natural human beings than other systems.**

***Index Terms*—Text-to-speech, Speech Synthesis, Prosody Modeling**

I. Introduction

Text-to-speech (TTS) has long been a focal point of research in the audio domain. Its goal is synthesizing speech from text with specified speaker timbre and style. During the synthesis, accurately modeling speech duration is crucial, especially for capturing natural prosody of human beings. Phoneme duration varies depending on the context and speaking style. Short phonemes may make the synthesized speech sound rushed, whereas elongated phonemes may emphasize certain words or enhance expressiveness. However, how to better model phoneme duration remains an open research question [1].

The two dominant TTS paradigms: autoregressive (AR) [2]–[4] and non-autoregressive (NAR) [5]–[7] systems, handle phoneme duration prediction differently. AR systems infer durations implicitly during generation, while NAR systems rely on explicit duration predictions for sequence alignment, ensuring that the Vocoder produces natural speech with prosodic variation. In NAR systems, this task is handled by a dedicated Duration Predictor (DP) module. Nevertheless, current NAR TTS systems struggle to accurately capture context-dependent phoneme duration variations, while also failing to model speaker-specific phoneme duration preferences.

Prior works try to analyze and improve DP module. [8] points out that whether using hard/soft alignment, or Monotonic Alignment Search [6] to estimate duration probability lower bounds, the resulting labels do not yield the ground truth. Overfitting to these labels leads to unreliable predictions, ultimately reducing the naturalness of synthesized speech. [1] further underscores the crucial role of DP and calls for more straightforward methods for prosody complexity assessment. [9] highlights the limitations of the popular NAR system VITS [10] in multi-speaker prosodic modeling and proposes using adversarial training to improve DP's accuracy. However, none of these approaches fundamentally restructured the DP module itself.

Meanwhile, our experiments reveal that more accurate and diverse prosodic information significantly increases the synthesis difficulty. When using HiFiGAN [11], [12], the dominant Vocoder in current NAR TTS [5], [6], [10], intricate prosodic details lead to substantial disharmony components on the spectrum. Similar issues have been identified in [13], [14], and they proposed their own alternative Vocoder frameworks. [15] further investigates the root cause of these disharmony components, attributing artifacts introduced by HiFiGAN's upsampling process and distortions from its downsampling process.

In this paper, we propose FNH-TTS, a novel end-to-end NAR TTS system based on VITS, with a particular focus on DP module design. To model prosodic diversity, we introduce the Mixture of Experts (MoE) [16], [17] mechanism. By employing a load-balancing loss, we encourage different experts to learn distinct prosodic patterns, enabling the DP module

*is the corresponding author.

to generate more precise and complex phoneme duration predictions. This, in turn, guides the Vocoder to synthesize speech that sounds more natural and human-like. To solve the synthesis problem raised from more diverse and complex prosodic information, we integrate an improved, faster Vocoder into FNH-TTS, paired with two advanced discriminators for adversarial training. This enables the Vocoder to handle complex prosodic variations, ensuring both naturalness and synthesis efficiency.

Our key contributions are as follows:

- We propose a novel Mixture of Experts Duration Predictor (MoE-DP), aiming to exploit its multi-expert nature to learn the diversity of prosody. This is the first instance of applying MoE to prosody modeling in TTS. Experimental results demonstrate a significant improvement in phoneme duration prediction accuracy.
- We integrate the VOCOS [14] Vocoder with Collaborative Multi-Band and Sub-Band Discriminators [15] into our end-to-end framework. These upgrades mitigate the disharmony components introduced by the increased prosodic complexity from the MoE-DP while maintaining inference efficiency.
- We introduce a novel visualization method for prosodic complexity analysis, offering an intuitive representation of how well the model captures speaker prosodic patterns. Additionally, we design a Vocoder evaluation scheme that circumvents length mismatches between target and synthesized audio, enabling more comprehensive comparisons of acoustic aspects. Lastly, we also reveal the unreliability of using Word Error Rate (WER) to assessing prosody of synthesized speech,

II. Methodology

Our system, FNH-TTS, retains the Text Encoder, Speaker Encoder, Posterior Encoder, and Flow modules from VITS, modifying the DP and Vocoder (Fig 1 left). This is made possible by the modular design philosophy of NAR TTS systems. The overall optimization objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{kl} + \mathcal{L}_{dur} + \mathcal{L}_{adv} + \mathcal{L}_{gen} \tag{1}$$

$\mathcal{L}_{rec}$ and $\mathcal{L}_{kl}$ are for the Text, Speaker, Posterior Encoder, and Flow. We retain these modules and loss functions without modification.

$\mathcal{L}_{dur}$ corresponds to the DP. We adopt the newly proposed Mixture-of-Experts Duration Predictor (MoE-DP). This module falls under the category of Deterministic Duration Predictors (DDP) [1], meaning it directly outputs phoneme durations. $\mathcal{L}_{dur}$ consists of two components: (1). $\mathcal{L}_{mas}$, computed using Monotonic Alignment Search (MAS) [6] result, (2). $\mathcal{L}_{aux}$, an auxiliary loss designed to balance the workload among different experts. Details are provided in Section II-A.

$\mathcal{L}_{gen}$ and $\mathcal{L}_{adv}$ are used for the Vocoder. To mitigate the synthesis challenges introduced by complex and diverse phoneme durations, we employ two advanced multi-scale discriminators to enhance the Vocoder's perception on both the temporal and frequency domain. The outputs of these discriminators are used to compute $\mathcal{L}_{adv}$. Additionally, to improve synthesis efficiency, we integrate the VOCOS vocoder into our end-to-end system. Details are provided in Section II-B.

A. MoE Duration Predictor

The Mixture of Experts (MoE) Routing mechanism, initially introduced by [16], aims to enhance model capacity without significantly increasing inference costs. Its architecture consists of a Router and a set of Experts. The Router Dynamically selects suitable Experts and distributes the input data accordingly, while each Expert specializes in processing specific kinds of inputs. The most popular MoE architecture at this stage is the Switch-Transformer [17]. It applies MoE on the Feed-Forward Network (FFN) to boost the transformer's capacity across diverse tasks without significantly increasing inference overhead.

Building on this concept, we propose a 1D convolution + Switch-Transformer framework for DP (Fig 1 middle). The input is $s+h_{text}$ from Speaker and Text Encoder, respectively. Meanwhile, we incorporate the s into the Router. The routing and expert processing mechanism is formulated as Eq 2. y represents the model output, Ψ denotes the top-k selected Experts from N based on p_i, x is the random variable from the previous layer, $p_i(x+s)$ is the weight assigned by the Router to Expert i, and $E_i(x)$ is the output of Expert i.

$$y = \sum_{i \in \Psi} p_i(x+s) E_i(x) \tag{2}$$

Meanwhile, to mitigate the unbalanced Expert assignment issue, we incorporate the Load Balancing Loss [17] as an auxiliary loss $\mathcal{L}_{aux}$ (Eq 3). The loss is scaled by α, while f_i (Eq 4) measures the proportion of tokens assigned to Expert i within a batch $\mathcal{B}$. Here, tokens refer to $x+s$. P_i (Eq 5) normalizes the probability of all tokens selecting Expert i across the batch $\mathcal{B}$, and T represents the total token set within the batch $\mathcal{B}$.

$$\mathcal{L}_{aux} = \alpha \cdot N \cdot \sum_{i=1}^{N} f_i \cdot P_i \tag{3}$$

$$f_i = \frac{1}{T} \sum_{token \in \mathcal{B}} \mathbb{1}\{\arg\max_{N} \ p(token) = i\} \tag{4}$$

$$P_i = \frac{1}{T} \sum_{token \in \mathcal{B}} p_i(token) \tag{5}$$

Finally, the overall optimization objective for the MoE-DP is defined as Eq 6.

$$\mathcal{L}_{dur} = \mathcal{L}_{mas} + \mathcal{L}_{aux} \tag{6}$$

B. Vocoder and Discriminators

Our experiments reveal that MoE-DP increases the difficulty of HifiGan synthesis, leading to a decline in synthesis quality (see Section IV-A). To address this, we employ VOCOS, a Vocoder framework based on ConvNeXt blocks [18] (Fig 1 right). VOCOS not only delivers superior synthesis quality but also offers accelerated inference speed.

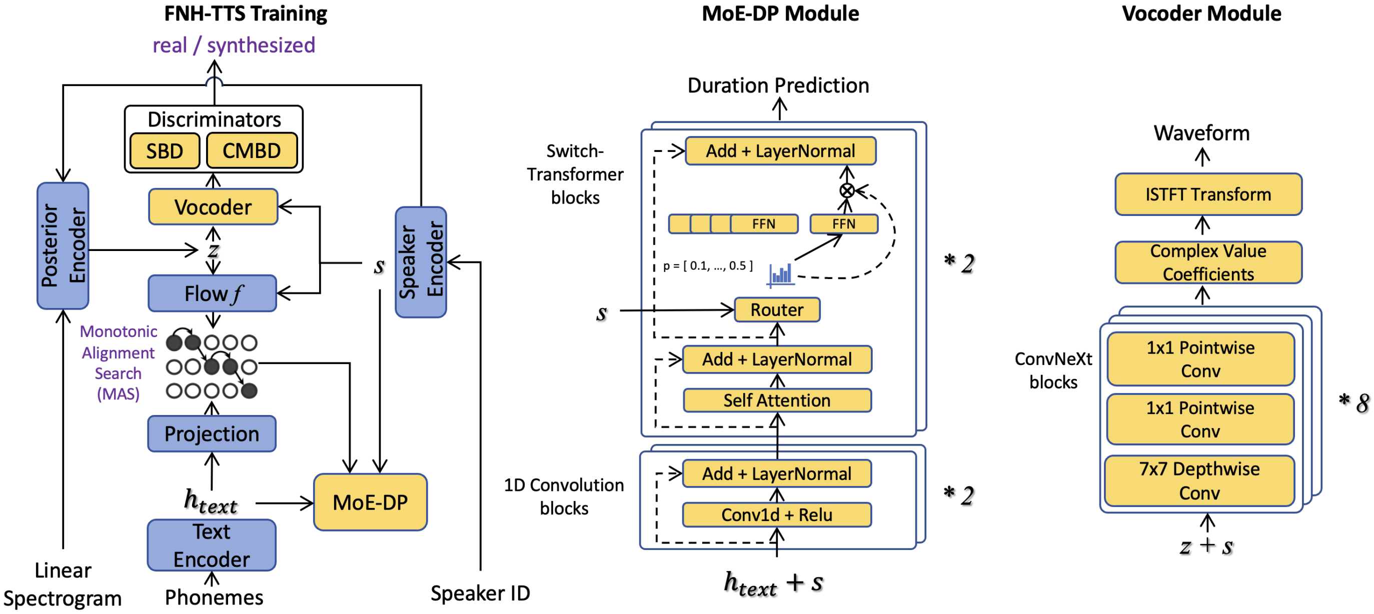

Fig. 1. The overall framework of the FNH-TTS system. The blue modules represent components inherited from the original VITS, while the yellow modules indicate newly introduced modifications.

To seamlessly integrate VOCOS into our end-to-end training pipeline, we replace its original Mel spectrum input with the latent variable z from the Posterior Encoder, along with the speaker embedding s (Eq 7). The Backbone network, built on ConvNeXt blocks, transforms the input features into a spectral representation, and ISTFT (Inverse Short-Time Fourier Transform) reconstructs the waveform $\hat{w}$ from the representation.

$$\hat{w} = ISTFT(Backbone(z + s)) \tag{7}$$

However, simply upgrading the Vocoder is insufficient to eliminate the disharmony components caused by increased prosodic diversity. Therefore, we adopt Collaborative Multi-Band Discriminator (CoMBD) and Sub-Band Discriminator (SBD) in our system, leveraging adversarial training to enhance temporal and spectral coherence. Our experiments confirm that these enhancements are crucial to FNH-TTS's success.

Collaborative Multi-Band Discriminator (CoMBD): This discriminator operates on waveform signals at different resolutions. To save computational resources, each waveform resolution is processed using the identical multi-scale discriminator (MSD) [19]. This approach ensures global waveform coherence while refining local waveform structures, leading to more natural temporal structures in the generated audio.

Sub-Band Discriminator (SBD): This discriminator employs PQMF analysis [20] to decompose the speech waveform into multiple sub-band signals, enabling precise discrimination across different frequency ranges. Each sub-band is applied to multi-scale dilated convolutions to extract features at varying receptive fields, facilitating more effective frequency-domain analysis. This discriminator is used for reducing high-frequency distortion and enhancing low-frequency stability, ensuring a more balanced synthesis quality across the spectrum.

C. *Model and Training Configuration*

For the Text, Speaker, Posterior Encoder, and Flow, we adopt the original VITS hyperparameters. For the MoE-DP, The convolution part consists of two 1D Convolution blocks with a kernel size of 3 and a dimension of 192. The MoE part comprises two Switch-Transformer blocks, each containing 8 experts, with 4 attention heads and a hidden dimension of 192. For the Vocoder, The hop length and FFT size are 256 and 1024. It contains 8 ConvNeXt blocks, each with an intermediate dimension of 1536 and a hidden dimension of 512.

Our system is trained using the AdamW optimizer, with $\beta_1 = 0.8, \beta_2 = 0.99$. The initial Learning Rate is 2*10-4, and a 0.999 decay factor is applied on every epoch. The batch size $\mathcal{B}$ is 24, and the training uses 4 NVIDIA RTX 3090 GPUs.

III. Evaluation

We systematically evaluate the contributions of different modules to the final synthesis quality using both single-speaker and multi-speaker TTS datasets: LJSpeech [21] and VCTK [22]. The test set consists of 1,324 samples held out from LJSpeech and 1,500 samples from VCTK. For synthesized quality evaluation, we adopt Mean Opinion Score (MOS) [23] and Word Error Rate (WER) as primary metrics. For MOS, we use a 5-point rating scale with 1-point intervals, and each sample is rated 10 times by 10 annotators. Only scores within the 95% confidence interval are retained to ensure reliability. For WER, we use the Whisper-Large-V3 [24] to transcribe the synthesized speech.

Meanwhile, to independently evaluate the Vocoder performance, we use the following metrics: Multi-Scale STFT Loss (M-STFT) [25], Perceptual Evaluation of Speech Quality (PESQ) [26], Mel Cepstral Distortion (MCD) [27], Periodicity, Voiced/Unvoiced F1 Score (V/UV F1) [28], plus the real-time

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT SYSTEMS.

System Description	LJSpeech MOS(↑)	LJSpeech WER(↓)	VCTK MOS(↑)	VCTK WER(↓)	Model Size	Discriminator Size
FastSpeech2	4.12	7.02%	-	-	34.64M	42.53M
StyleTTS2	4.35	5.78%	-	-	145.53M	41.38M
F5-TTS	3.87	8.70%	4.49	**2.24%**	337.09M	-
SparkTTS	4.10	7.36%	4.38	4.98%	506.63M	-
VITS Origin	4.26	3.41%	4.34	4.11%	39.53M	46.75M
| - w/ CMoBD + SBD	4.32	2.69%	4.50	4.00%	39.53M	27.07M
| ∟ w/ VOCOS	4.44	2.60%	4.27	4.96%	40.89M	27.07M
∟ - w/ MOE-DP	3.92	6.48%	3.90	7.04%	46.37M	46.75M
∟ w/ VOCOS	3.75	9.74%	3.95	10.58%	47.73M	46.75M
∟ w/ CMoBD + SBD	4.20	**2.42%**	4.43	3.60%	46.37M	27.07M
∟ w/ VOCOS (**our FNH-TTS**)	**4.48**	2.59%	**4.63**	3.88%	47.73M	27.07M

TABLE II
COMPARISON RESULTS OF DIFFERENT VOCODERS

Systems	M-STFT(↓)	PESQ(↑)	MCD(↓)	Periodicity(↓)	V/UV F1(↑)	RTF(CPU)(↓)	RTF(GPU)(↓)
VITS(HifiGan)	1.208	**2.441**	1.677	**0.160**	**0.928**	0.352	0.00748
VITS(HifiGanV2)	1.246	2.231	1.814	0.166	0.922	0.181	0.00659
FNH-TTS	**1.207**	2.425	**1.654**	0.162	0.927	**0.046**	**0.00460**

factor (RTF) results for inference speed. Since synthesized and original audio often differ in length, end-to-end TTS systems cannot directly compute these metrics. Therefore, we only retain the Posterior Encoder and Vocoder for evaluation. We first extract the linear spectrogram from the audio, then encode it via Posterior Encoder to obtain latent variable z. In the end, we reconstruct the audio from z via Vocoder. This encoding-decoding process does not involve DP, thus ensuring length consistency while also demonstrating the Vocoder performance.

Finally, to assess phoneme duration prediction accuracy, we conduct experiments on the LibriTTS [29] 100+360 (Libri460) dataset, analyzing the accuracy and distribution differences of the predictions across systems. Phoneme duration annotations are sourced from [30], and we extract 111 test samples for evaluation. The evaluation method is as follows: Using Montreal Forced Alignment (MFA) [31], we perform forced alignment on both synthesized and ground truth (GT) audio, calculating average phone duration and mapping it into three prosody categories: high, normal, and low. We then compute accuracy and visualize distribution fitting curves to compare each system's ability to model speaker prosody.

IV. RESULTS

In this section, we first compare our system with other TTS systems to demonstrate the effectiveness of FNH-TTS. Then, we present a comparative analysis of the vocoder and DP module to provide preliminary evidence of their contributions to the overall performance of the system.

A. Synthesis Quality Comparison

Table I presents the comparison results between our systems and other SOTA TTS systems, as well as the contribution of each proposed module to overall performance. The key observations are as follows:

- FNH-TTS achieves the highest MOS score, outperforming all other systems. In terms of WER, the configuration with MoE-DP + CMoBD + SBD achieves the best performance in single-speaker scenarios (LJSpeech). Our system falls slightly behind F5-TTS in multi-speaker settings (VCTK). However, considering model size and deployment feasibility, especially for on-device applications, FNH-TTS holds a clear advantage.
- The best-performing systems require both MoE-DP and the CMoBD + SBD. VOCOS yields limited improvements in MOS and WER metrics. Nevertheless, due to VOCOS's clear advantages in efficiency, acoustic metrics (see Section IV-B), and spectrogram quality (see Section V-B), we designate the system incorporating all proposed enhancements as FNH-TTS.
- Introducing MoE-DP alone degrades synthesis quality, and its full potential is only unleashed when combined with upgraded discriminators. This is because MoE-DP introduces more diverse prosody predictions, which increases the difficulty of vocoder synthesis. For the comparison of different systems on prosody diversity and prediction accuracy, see Section IV-C. For the detailed discussion of the roles of MoE-DP, VOCOS, and the discriminators, see Section V.

B. Vocoder Performance Evaluation

We compare FNH-TTS against the original VITS (with HiFiGAN [11]) and its improved variant (using HiFiGANv2 [12]). The RTF on both CPU and GPU is also measured. The evaluation procedure follows the approach outlined in Section III to address length mismatch between synthesis and reference audio. We use Intel Xeon Gold 5218R @ 2.10GHz and NVIDIA RTX 3090 for CPU and GPU inference, respectively.

All results are in Table II. It shows that our system's Vocoder outperforms all other systems regarding spectral accuracy (M-STFT) and timbre fidelity (MCD). However, it slightly lags behind in artifact reduction, naturalness, and pronunciation accuracy (PESQ, Periodicity, and V/UV F1). Despite incorporating CMoBD and SBD, the stronger prosodic information introduced by the MoE-DP still pose a huge chal-

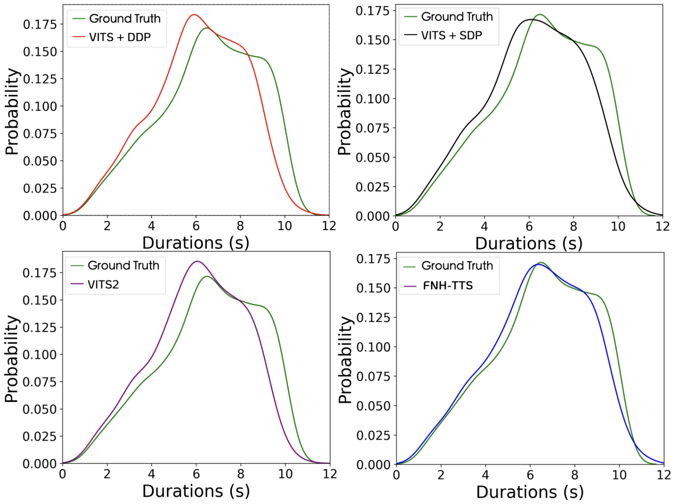

Fig. 2. Phone duration distributions of different systems on LJSpeech.

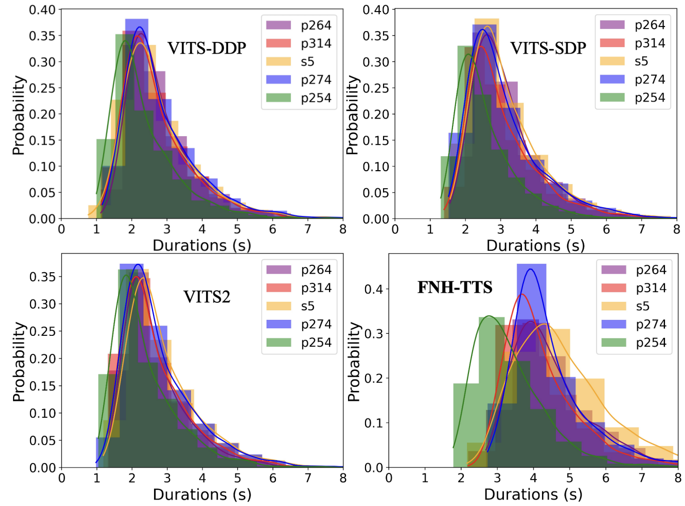

Fig. 3. Phone duration distributions of different systems on VCTK.

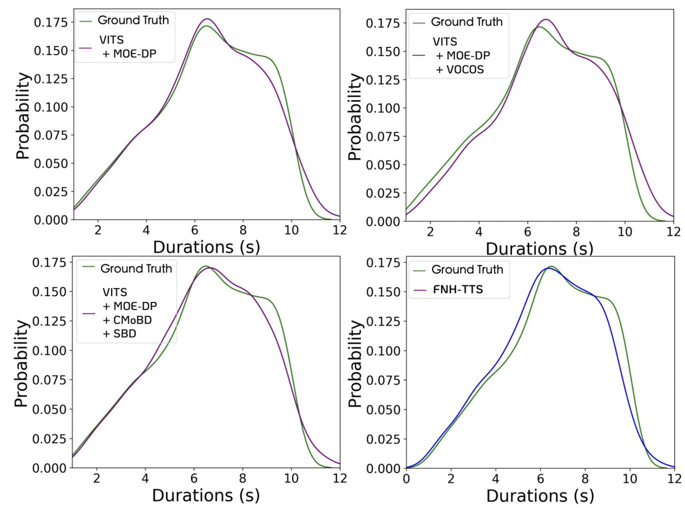

Fig. 4. Phone duration distributions of using different settings on LJSpeech.

lenge for Vocoder, leading to disharmony components such as artifacts and distortions. Regarding inference speed, our Vocoder significantly outperforms all other systems, demonstrating superior efficiency in synthesis.

C. Prosody Modeling Analysis

TABLE III
COMPARISON OF DIFFERENT SYSTEMS ON LIBRI460.

Systems	Libri460 ACC(↑)	Libri460 MOS(↑)	Libri460 WER(↓)
FastSpeech2	59.80%	4.37	2.12%
StyleTTS2	59.28%	4.41	**1.58%**
F5-TTS	11.17%	4.35	4.21%
SparkTTS	66.77%	4.36	5.12%
VITS+DDP	61.26%	4.37	3.42%
VITS+SDP	65.76%	4.27	6.06%
VITS2	60.36%	4.33	3.34%
FNH-TTS	**67.07%**	**4.42**	3.15%

We compare FNH-TTS with other SOTA TTS systems in terms of phone duration prediction accuracy, MOS, and WER on Libri460. Additionally, we include comparisons with the original VITS (with DDP or the stochastic duration predictor, SDP [1], [10]) and VITS2 (which specifically targets improvements in phone duration prediction), as shown in Table III. The key observations are as follows:

- FNH-TTS achieves the highest accuracy among all evaluated systems. In terms of MOS and WER, FNH-TTS also outperforms most baseline systems, with the sole exception of StyleTTS2.
- Although WER is widely used to evaluate TTS synthesis quality, it is not an ideal metric for capturing prosody modeling performance. In fact, weaker prosody modeling can sometimes result in lower WER, as ASR systems are less disrupted by complex or natural prosodic variations.

To further investigate its underlying mechanism, we visualized the prediction distributions for different systems' DPs on LJSpeech (single-speaker dataset) (Fig 2) and VCTK (multi-speaker dataset)(Fig 3) test sets. For VCTK, we highlight the prediction distributions of five randomly selected speakers (p264, p314, s5, p274, and p254) using different colors to illustrate speaker-specific variations. Here are our findings:

- On single-speaker data, FNH-TTS's prediction distribution aligns more closely with the ground truth. VITS-DDP and VITS2 show a sharp peak, and VITS-SDP follows a more Gaussian-like distribution. These competitors' distribution deviates from natural prosodic patterns of speech.
- On multi-speaker data, FNH-TTS demonstrates stronger speaker distinguish ability. It captures distinct prosodic characteristics for different speakers. In contrast, other models predict nearly identical distributions across different speakers, which contradicts the natural variation in prosody among different individuals.

V. DISCUSSION

In Section IV-A, our experiments reveal that using MOE-DP alone negatively impacts the final synthesis quality of the TTS system. Its benefits can only be effectively unlocked when combined with VOCOS and the CMoBD+SBD. Therefore, we further analyze the role of MOE-DP in prosody modeling and

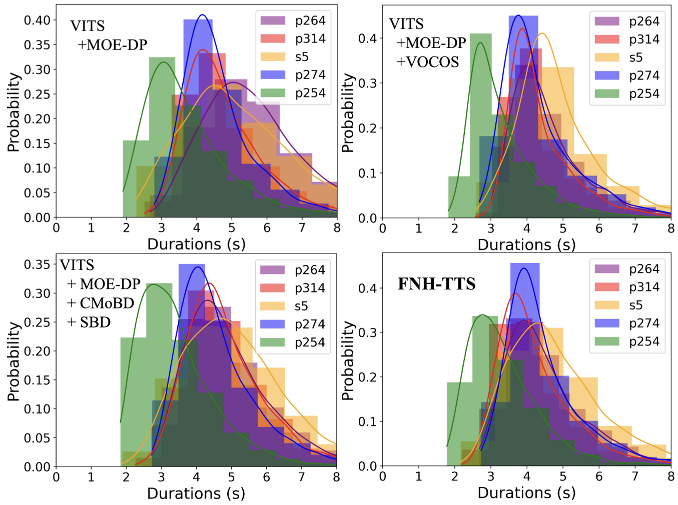

Fig. 5. Phone duration distributions of using different settings on VCTK.

TABLE IV
THE JENSEN–SHANNON DIVERGENCE OF THE DP OUTPUT BETWEEN FNH-TTS AND OTHER SYSTEMS

System Description	JS divergence			
	LJSpeech		VCTK	
	Mean	Var	Mean	Var
VITS+SDP	0.087	0.0003	0.066	0.0001
VITS+DDP	0.057	0.0002	0.043	0.0002
VITS+MOE-DP	0.053	0.0002	0.039	0.0001
VITS+MOE-DP+VOCOS	0.053	0.0002	0.045	0.0001
VITS+MOE-DP+CMoBD+SBD	0.047	0.0002	0.041	0.0004

the role of the decoder framework in handling more complex prosodic diversity.

A. The Role of MOE-DP in Prosody Modeling

To demonstrate MOE-DP's contribution to prosody modeling, we visualize duration predictions across several configurations: VITS+MOE-DP, VITS+MOE-DP+VOCOS, VITS+MOE+CMoBD+SBD, and FNH-TTS on both the LJSpeech and VCTK datasets (Figures 4 and 5). As shown, the inclusion of MOE-DP alone has already yielded a closer alignment between prediction and ground truth in single-speaker scenario. In multi-speaker scenarios, it also captures prosodic diversity effectively as long as MOE-DP is involved in the system.

To further validate this observation, we use the DP output of FNH-TTS as a baseline and quantify the divergence between this baseline and the outputs of TTS systems integrated with different modules. Specifically, we compute the Jensen–Shannon (JS) divergence for each sample in the LJSpeech and VCTK test sets, followed by calculating the mean and variance across the whole dataset. The results, shown in Table IV, indicate that in terms of phoneme duration divergence, the primary contributors are MOE-DP and the two new discriminators. In single-speaker scenarios (LJSpeech), their contributions are nearly equivalent, while in multi-speaker settings (VCTK), MOE-DP demonstrates a significantly greater impact compared to the discriminators.

B. Advantages of VOCOS and CMoBD+SBD in Handling Prosodic Diversity

However, Section V-A observations raises a more intriguing question: why does using MOE-DP alone benefit prosody modeling but fail to improve the final synthesis quality? To investigate the underlying reasons, we visualize the spectrograms of the following systems: VITS + MOE-DP, VITS + MOE-DP + VOCOS, VITS + MOE-DP + CMoBD+SBD, and FNH-TTS (Figure 6).

From the Figure 6, we can clearly draw three conclusions that highlight how VOCOS and CMoBD+SBD enhance prosody modeling:

- **With MOE-DP alone**, the generated waveform exhibits noticeable truncation and lacks overall continuity (see region 2 in the Figure 6). Additionally, the texture sharpness is suboptimal, especially in the high- and low-frequency bands (see region 1), where the features appear significantly weaker compared to the other systems. This suggests that the conventional HiFi-GAN combined with the multi-period discriminator fails to adequately handle the nuanced diversity of prosodic patterns, leading to rough and less natural synthesis.
- **The inclusion of VOCOS** significantly sharpens the waveform texture, particularly in the high- and low-frequency regions (see region 1), indicating an improvement in fine-grained signal synthesis.
- **Adding CMoBD and SBD** substantially improves the continuity and smoothness of the waveform texture. As seen in region 2, these discriminators enhance both the head and tail segments, producing clearer and more fluid transitions. Even under large prosodic variations, the waveform structure remains consistent and coherent.

VI. CONCLUSION

We propose FNH-TTS, a faster, more natural, and Human-Like TTS system that better aligns with human speech prosodic patterns. Our approach introduces the MoE-DP module, leveraging the MoE mechanism to capture speaker-specific prosodic variations more effectively. To mitigate the increased synthesis complexity caused by diverse prosodic patterns, we upgrade Vocoder and discriminators, addressing unnatural components that arise from richer and more complex prosodic information. Our experiments demonstrate the effectiveness of FNH-TTS and also reveal the broader impact of complex prosodic information on modern TTS systems. Through visualization analysis, we highlight FNH-TTS's significant improvements in prosody imitation. However, both experimental and visual results indicate that current vocoders still have room for improvement in handling intricate prosodic patterns. Smoothing the synthesis under diverse prosodic pattern remains a major challenge for future research. Finally, we analyze the contributions of individual components to the FNH-TTS and point out the limitations of current evaluation metrics, such as WER, in assessing prosody. We hope our work will further advance research in prosody modeling and contribute to building more human-like speech synthesis systems.

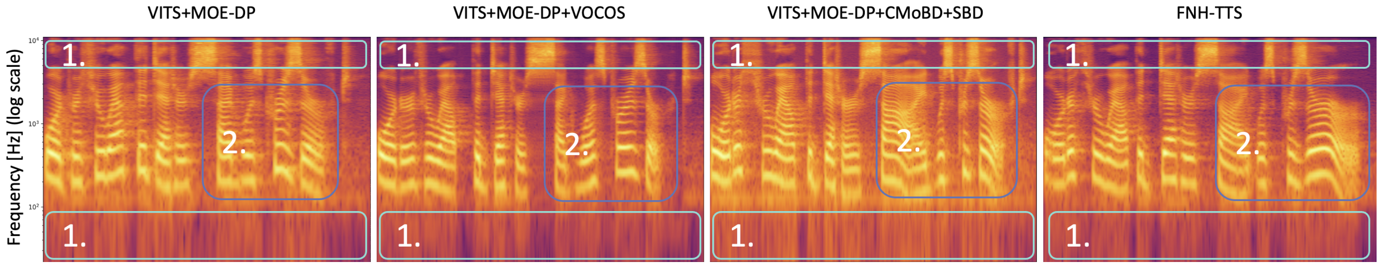

Fig. 6. Four spectrograms of LJSpeech sample audio from VITS+MOE-DP, VITS+MOE-DP+VOCOS, VITS+MOE-DP+CMoBD+SBD, and the FNH-TTS, respectively.

REFERENCES

[1] S. Mehta, H. Lameris, R. Punmiya, J. Beskow, É. Székely, and G. E. Henter, "Should you use a probabilistic duration model in tts? probably! especially for spontaneous speech," *arXiv preprint arXiv:2406.05401*, 2024.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[4] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers," *arXiv preprint arXiv:2406.05370*, 2024.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[6] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[7] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.

[8] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, 2025.

[9] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," *arXiv preprint arXiv:2307.16430*, 2023.

[10] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

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

[12] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 166–170.

[13] M. Kawamura, Y. Shirahata, R. Yamamoto, and K. Tachibana, "Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," *arXiv preprint arXiv:2306.00814*, 2023.

[15] T. Bak, J. Lee, H. Bae, J. Yang, J.-S. Bae, and Y.-S. Joo, "Avocodo: Generative adversarial network for artifact-free vocoder," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 562–12 570.

[16] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "The sparsely-gated mixture-of-experts layer," *Outrageously large neural networks*, 2017.

[17] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

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

[20] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-qmf banks," *IEEE Transactions on signal processing*, vol. 42, no. 1, pp. 65–76, 1994.

[21] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[22] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[23] M. Chu and H. Peng, "An objective measure for estimating mos of synthesized speech." in *INTERSPEECH*, 2001, pp. 2087–2090.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[25] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.

[28] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive gan for conditional waveform synthesis," *arXiv preprint arXiv:2110.10139*, 2021.

[29] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[30] S. Ji, J. Zuo, M. Fang, Z. Jiang, F. Chen, X. Duan, B. Huai, and Z. Zhao, "Textrolspeech: A text style control speech corpus with codec language text-to-speech models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 301–10 305.

[31] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.