Supplementary Material for P-Flow

Anonymous Author(s)
Affiliation
Address
email

1 Demo Page Link

The link to our demo page is https://bit.ly/3ID5Zam.

2 Additional Results

2.1 Effect of Euler Steps for Acoustic Quality

We present the objective metrics according to the Euler steps in the result section of the main paper. Since these objective metrics have limitations in representing acoustic quality with respect to the Euler step, we also evaluate the sample quality based on the Euler steps and provide it as an additional metric. We measure the acoustic quality using 5-scale Mean Opinion Scores (MOS). We inquire each human evaluator to assess the acoustic quality of each sample, and we evaluate it with the participation of more than 50 human evaluators.

Table 1 presents MOS along with SECS and inference latency shown in the results section, based on the Euler step N. The table demonstrates that as the number of Euler steps increases, the acoustic sample quality improves. We choose Euler step 10 as the default, as it ensures a high speaker similarity while providing a good balance between inference latency and sample quality.

Table 1: Mean Opinion Scores (MOS) for the acoustic quality and Objective Metrics according to the Euler steps N.

MODEL	N	MOS↑	SECS	INFERENCE LATENCY(S)↓
P-FLOW	1	3.55 ± 0.16	0.420	0.028 ± 0.004
	2	3.71 ± 0.12	0.522	0.037 ± 0.004
	5	4.01 ± 0.10	0.549	0.067 ± 0.004
	10	4.08 ± 0.10	0.544	0.115 ± 0.004
	20	4.14 ± 0.10	0.540	0.210 ± 0.005

2.2 Zero-shot TTS with Emotional Reference Speech

We provide generated samples using emotional reference samples, where each sample exhibits distinct prosody, as demonstrated in [4]. We extract reference speech samples from EmoV-DB [1], representing five different emotions. From each reference speech, we utilize a 3-second segment to perform zero-shot TTS. On our demo page, we present generated samples for the same sentence given the speech prompts for these five emotions. P-Flow, similar to VALL-E, utilizes a speech-prompted text encoder composed of an autoregressive transformer, enabling the generation of samples with different prosody based on the reference speech.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

3 Model Architectures

We provide explanations for each module in this section and detailed hyperparameters and architecture of P-Flow are shown in Table 2.

Speech-prompted Text Encoder Our text-encoder consists of several linear projection layers, a pre-network with 3 convolutional layers, and a 6-layer transformer with 2 attention heads of 192 hidden dimensions. The input to the text encoder is the speech prompt and text embeddings projected into the same dimensions. For the input of the speech-prompted text encoder, we project the speech prompt and text embeddings into the same dimension and input to the same pre-network. The resulting representation is then split into prompt and text parts, to which positional encodings are added. We define each positional encoding as the sum of absolute positional encoding and a learnable fixed-size embedding so that the transformer can differentiate the speech prompt and text through learnable embeddings. The representations of the speech prompt and text are then fed into a transformer architecture that allows each text position to attend to the speech prompt.

Duration predictor Our duration predictor is a shallow convolution-based model used in [2]. Since our text encoder output already provides speaker-conditional hidden representation, we use the hidden representation before linear projection to h_c as the input of the duration predictor.

Flow matching Decoder Our flow matching decoder utilizes 18 layers of WaveNet-like architecture [3] with 512 hidden dimensions. We use the global conditioning method in WaveNet for conditioning t and concatenate the aligned encoder output h with the input x_t along the channel axis for conditioning the speaker-conditional text representation.

Table 2: Hyperparameters of P-Flow

	Hyperparameter	
Speech-prompted Text Encoder	Phoneme Embedding Dim	192
	PreNet Conv Layers	3
	PreNet Hidden Dim	192
	PreNet Kernel Size	5
	PreNet Dropout	0.5
	Transformer Layers	6
	Transformer Hidden Dim	192
	Transformer Feed-forward Hidden Dim	768
	Transformer Attention Heads	2
	Transformer Dropout	0.1
	Prompt Embedding Dim	192
	Number of Parameters	3.37M
Duration Predictor	Conv Layers	3
	Conv Hidden Dim	256
	LayerNorm Layers	2
	Dropout	0.1
	Number of Parameters	0.36M
Flow Matching Decoder	WaveNet Residual Channel Size	512
	WaveNet Residual Blocks	18
	WaveNet Dilated Layers	3
	WaveNet Dilation Rate	2
	Number of Parameters	40.68M

References

[1] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.

[2] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Advances in Neural Information Processing Systems*, 33, 2020.

[3] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[4] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.