VarianceFlow: High-quality and Controllable Text-to-Speech Using Variance Information via Normalizing Flow

Yoonhyung Lee[1], Jinhyeok Yang[2], Kyomin Jung[1]

[1]Seoul National University, Dept. of Electrical and Computer Engineering, [2]Speech AI Lab, NCSOFT

Overview

- We propose a **non-autoregressive Text-to-Speech** model called **VarianceFlow**, which **takes variance information** such as pitch or energy as additional input during training.
- We suggest a new method to feed the variance information through a **Normalizing Flow** (NF) module rather than directly, where the module performs **modeling of the variance distribution.**
- By performing the variance modeling based on NF, we improve the **speech quality** and **variance controllability** of VarianceFlow.
- In experiments, VarianceFlow outperforms the previous SOTA AR and non-AR TTS models in terms of speech quality.
- In addition, it provides a more accurate control over the variance information compared to the widely-used baseline non-AR TTS model, FastSpeech 2.

Background

One-to-many problem in Text-to-Speech

- When modeling TTS, **one-to-many problem** should be considered for better performance (i.e. there are many ways to pronounce a single sentence).
- For **AR TTS models**, however, the one-to-many problem is naturally **resolved to some degree**, because it normally learns to generate a melspectrogram frame given the previous frames as well as the text.
- However, **AR TTS models have inevitable problems**: (1) Slow inference speed; (2) Error vulnerability. Therefore, non-AR TTS models recently have been proposed.

Two types of solutions for Non-AR TTS models to solve the one-to-many problem

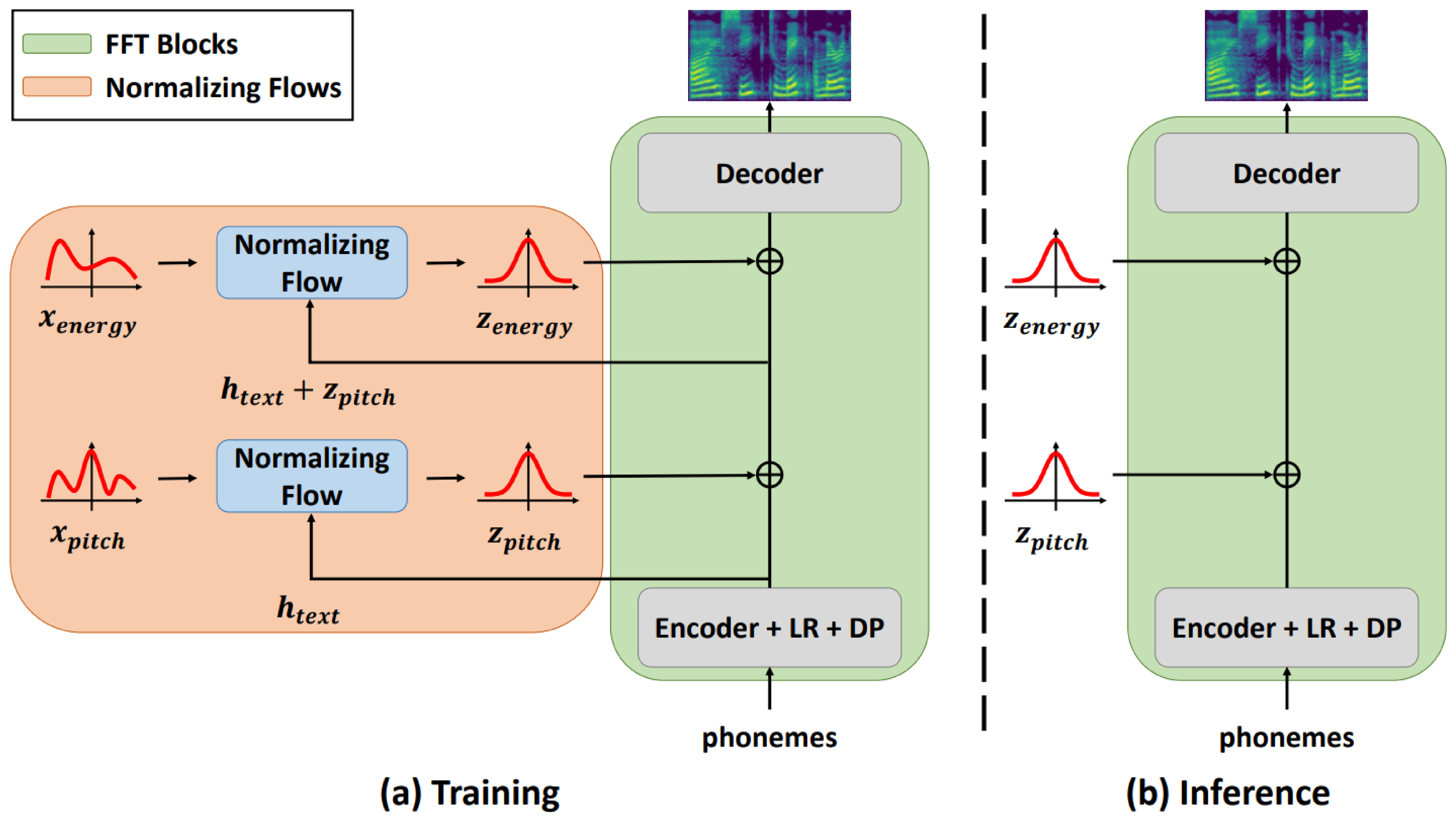
- **Type Ⅰ:** adopting **more flexible generative frameworks** such as Normalizing Flow or Score-based models (i.e. MSE-based training assumes the Gaussian distribution).
  ex) Glow-TTS [1], Grad-TTS [2]
- **Type Ⅱ:** explicitly **using variance information such as pitch or energy during training**, which significantly eases the one-to-many problem. It also allows models to explicitly control the variance information.
  ex) FastSpeech 2 [3], FastPitch [4]

⇒ We solve the problem remaining in FastSpeech 2 (Type Ⅱ) by adopting the idea used in Type Ⅰ models.

FastSpeech 2

- During training, **FastSpeech 2 directly takes the variance information** such as pitch or energy as well as a text input.
- Meanwhile, it has a module called **variance predictor**, which is jointly **trained to predict the variance information** from the text input based on **MSE loss.**
- At inference, FastSpeech 2 first **predicts the variance information** based on the input text using its variance predictor, and then it **generates speech using the predicted variance values** and text representations.
- **However, one-to-many problem also exists in predicting the variance information from the text input.**

VarianceFlow



- Unlike FastSpeech 2, **VarianceFlow takes variance information through a NF module**, which performs modeling of the variance information.
- At inference, it uses latent representations for the variance information by directly sampling them from simple prior distributions. (e.g. Gaussian distribution)
- Due to the flexibility of NF compared to MSE-based training, **it performs more accurate distribution modeling** resulting in improved speech quality.
- In addition, the training principle of **NF disentangles the text input and variance information**, which results in better controllability of the variance information.

Experiments and Results

Speech quality comparison

Table 1. MOS results written with 95% confidence intervals.

Model	MOS
GT Waveform	4.47 ± 0.07
GT Melspectrogram	4.34 ± 0.08
Tacotron 2	4.03 ± 0.07
Glow-TTS	3.72 ± 0.13
FastSpeech 2-phoneme	3.92 ± 0.07
FastSpeech 2-frame	3.66 ± 0.09
VarianceFlow-phoneme	4.04 ± 0.08
VarianceFlow-frame	**4.19 ± 0.07**

- In terms of speech quality, **VarianceFlow outperforms the previous SOTA AR and non-AR TTS models**, Tacotron 2, Glow-TTS, and FastSpeech 2.
- Also, we observe that the improvement in variance modeling performance is reflected in the results, where **only VarianceFlow benefits from performing finer variance modeling.**

Variance controllability comparison

Table 2. FFE (%) and MOS (score 1-5, 9-scale) results measured with different pitch shift scale λ.

Model	$\lambda=-4$		$\lambda=-2$		$\lambda=+2$		$\lambda=+4$	
	FFE	MOS	FFE	MOS	FFE	MOS	FFE	MOS
FastSpeech 2	14.00	3.46	12.61	3.65	10.94	3.29	11.57	2.63
VarianceFlow-reversed	35.97	4.01	53.47	4.00	66.37	3.90	67.07	3.69
VarianceFlow	12.16	3.87	9.02	4.05	7.26	3.95	7.52	3.39

- While varying pitch input by multiplying a positive scalar to the pitch values, we measure MOS and f0 frame error rates between the pitch input and the pitch calculated from generated speech.
- Here, **VarianceFlow shows lower FFE while maintaining better speech quality.**
- Also, using the **variance information through a NF shows its effectiveness in disentagleing** the text and variance information.

References

[1] Kim, et al, “Glow-tts: A generative flow for text-tospeech via monotonic alignment search,” in Proc. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 8067–8077.

[2] Popov, et al., “Grad-tts: A diffusion probabilistic model for text-to-speech,” in Proc. Int. Conf. on Machine Learning (ICML), 2021, vol. 139, pp. 8599–8608.

[3] Ren, et al., “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in Proc. Int. Conf. on Learning Representations (ICLR). 2021, OpenReview.net.

[4] Adrian Lancucki, “Fastpitch: Parallel text-to-speech with pitch prediction,” in Proc. ICASSP, 2021, pp. 6588–6592.