#2896

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

Yoonhyung Lee, Jinhyeok Yang, Kyomin Jung

Machine Intelligence Lab, Seoul National University

Speech AI Lab, NCSOFT

ICASSP 2022

Authors

Yoonhyung Lee

Seoul National Univ.

Jinhyeok Yang

NCSOFT

Kyomin Jung

Seoul National Univ.

Overview

- We propose **a non-autoregressive TTS 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 distribution modeling of the variance information.

- By performing the variance modeling based on NF, **we improve the speech quality and variance controllability** of VarianceFlow.

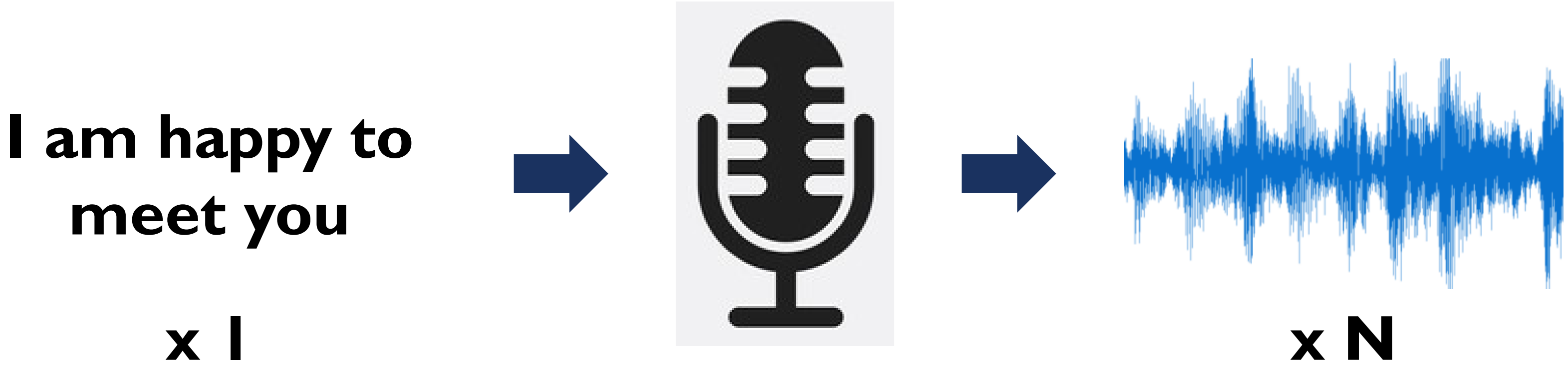
Table of Contents

1. One-to-many problem in TTS
2. VarianceFlow
3. Experiments
4. Conclusion

One-to-many relationship in TTS

One-to-many relationship in TTS

- There are various ways to read a given sentence, even if it’s the same speaker.
  (e.g. different speed, different pitch, different volume…)



I am happy to meet you

x 1

x N

Autoregressive TTS

- Autoregressive Text-to-Speech model

$$\boxed{p(\underline{x_t}|x_{<t},\underline{Y})}$$

't'-th frame of a melspectrogram

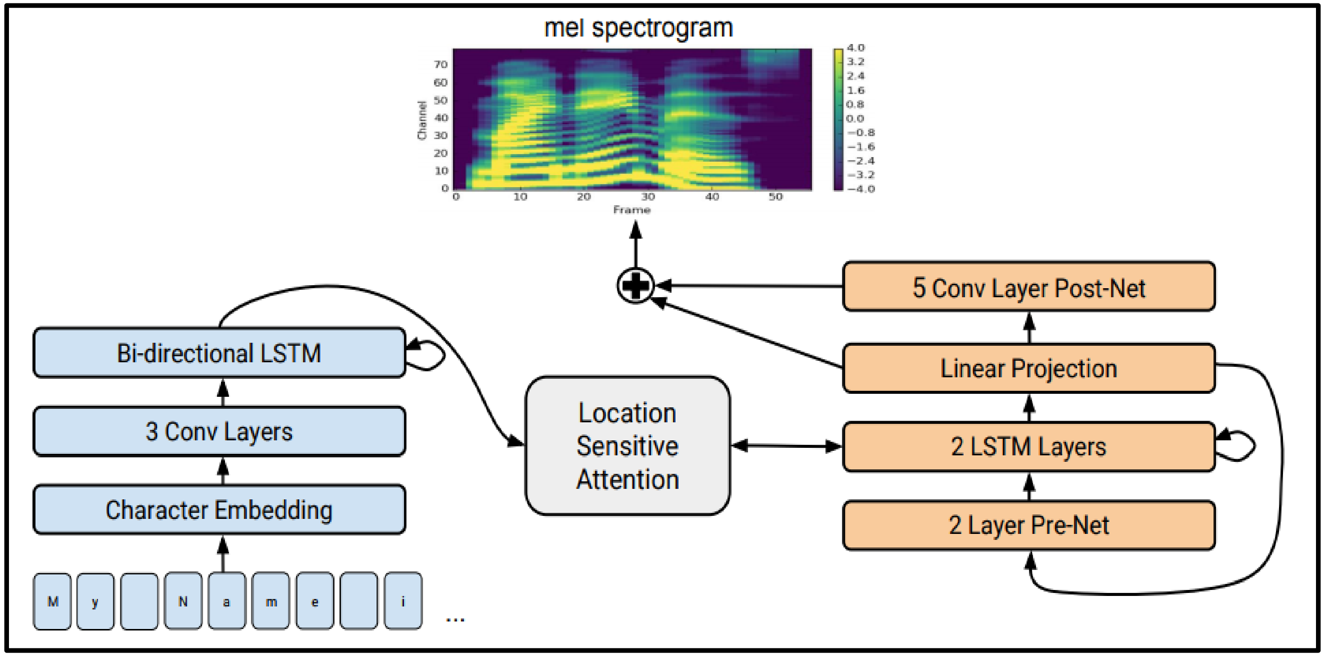
text input (phoneme sequence)

- It is trained to generate a melspectrogram frame based on the previous melspectrogram frames and text input.

- The one-to-many property is less problematic for autoregressive modeling.

Autoregressive TTS

〈Tacotron 2〉



- It generates a melspectrogram frame-by-frame, so it is slow.
- It is vulnerable to attention errors and errors occurred due to the exposure bias.

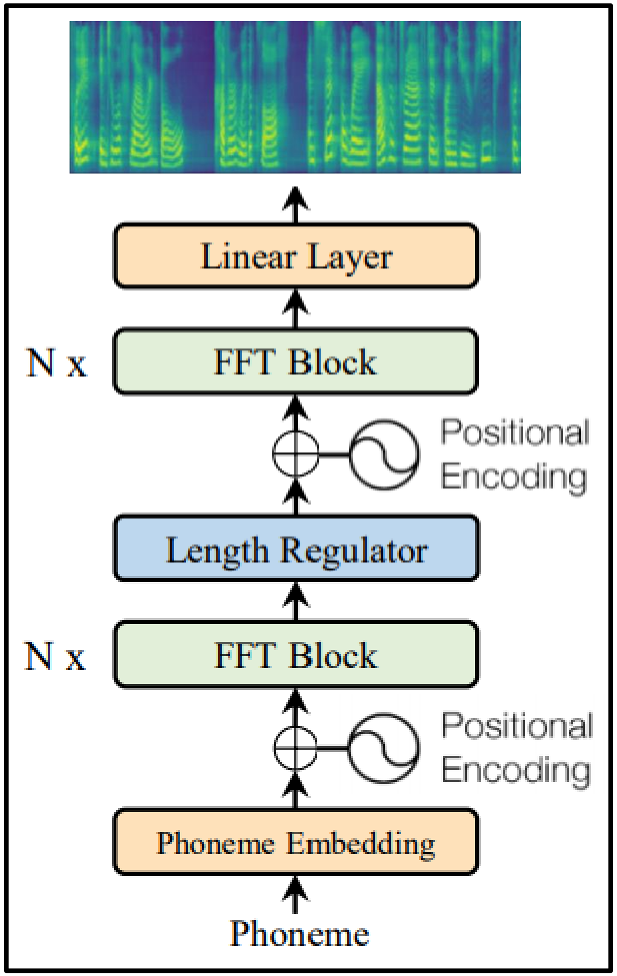
Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.

Non-AR TTS

<FastSpeech>

- Using alignment information, it is trained to generate melspectrograms with mean absolute error (MAE).

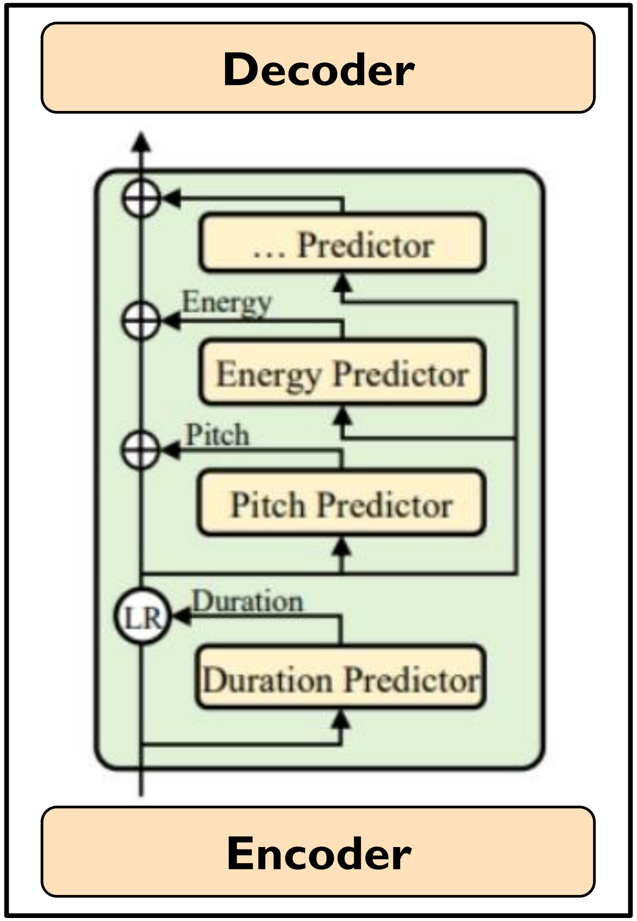
- Using MAE means it is assuming the Laplacian target distribution.
  → contradict to the one-to-many property, resulting in bad speech quality



Yi Ren et al., Fastspeech: Fast, robust and controllable text to speech. In Advances in Neural Information Processing Systems, pp. 3171–3180, 2019.

Non-AR TTS

〈FastSpeech 2〉



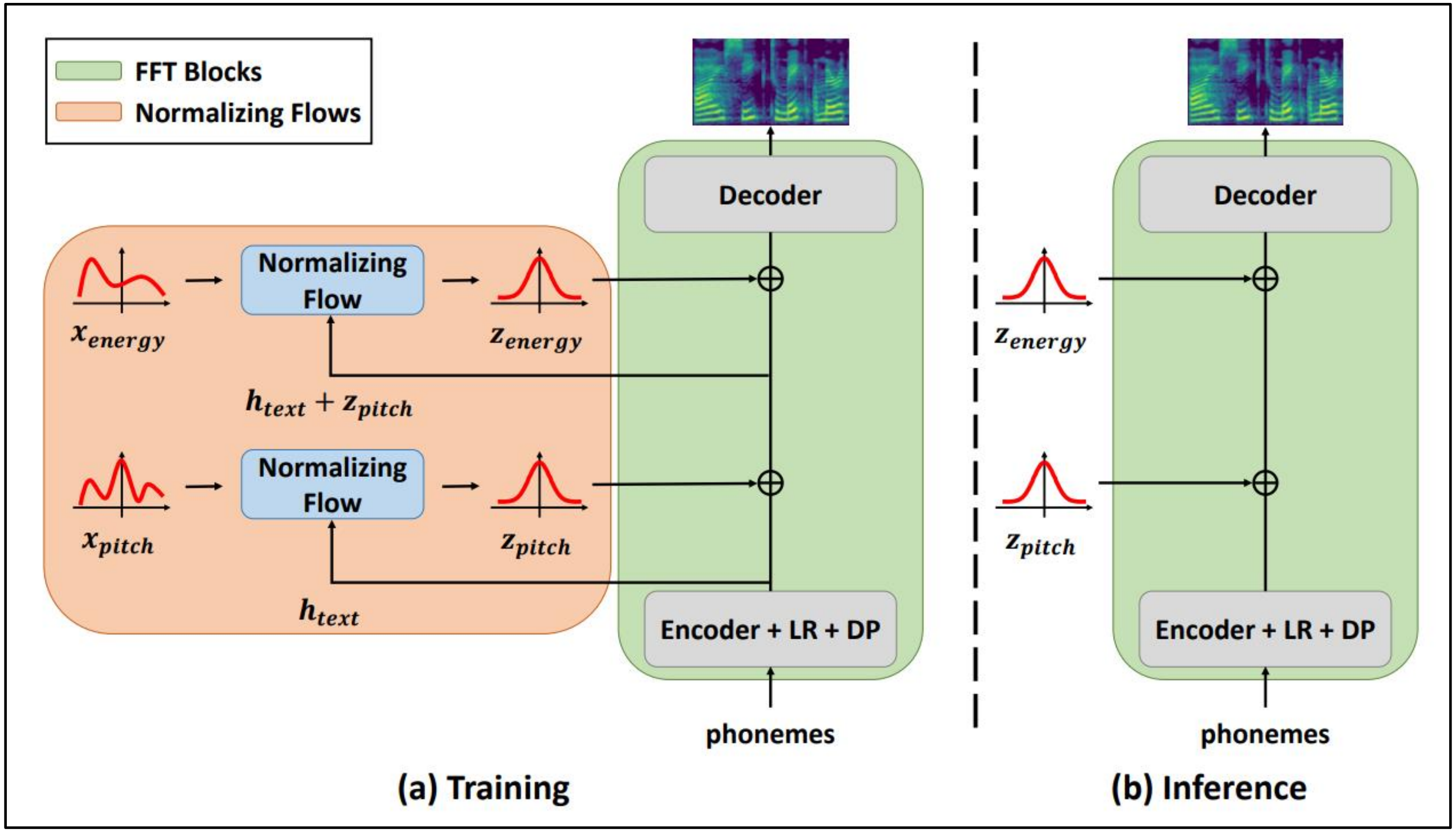
- Using variance information, it alleviates the one-to-many problem.
- Also, it allows to control the variance factors
- Variance predictors are trained to predict the variance values based on text preparing for inference.
- Variance prediction is trained with L2-loss, however, there exists also the one-to-many relationship.

Yi Ren et al., FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, International Conference on Learning Representations (ICLR), 2021.

VarianceFlow

VarianceFlow

- We replace the variance predictors of FS2 with NF modules, which perform modeling of the variance distribution better.



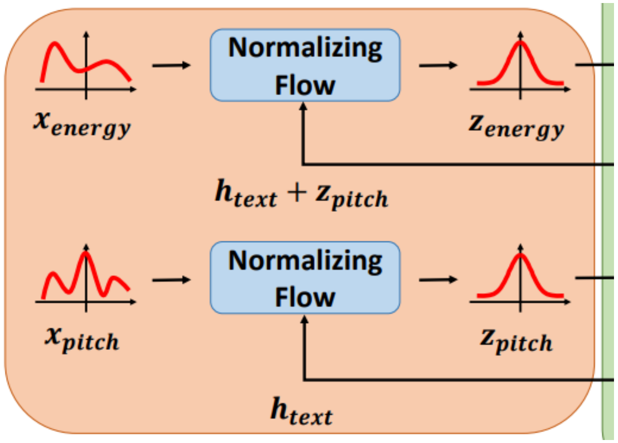
VarianceFlow

- During training, NF modules are trained to minimize the below loss based on invertible transforms.

$$\boxed{\mathcal{L}_{NF} = D_{KL}\left[q_\theta(z|h) \parallel p(z)\right] + H(x|h)}$$

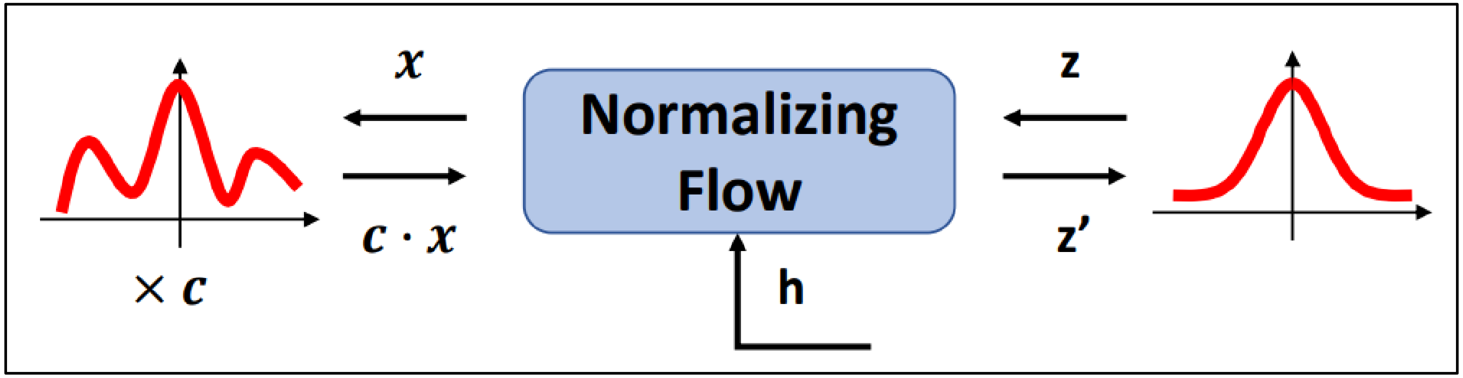
- NF does not assume pre-defined distribution shape.
  → improvement in variance modeling performance

- Indirectly feeding the variance information leads to the disentanglement of h and z.
  → improvement in variance controllability



VarianceFlow

- Variance Control using the invertibility of NF

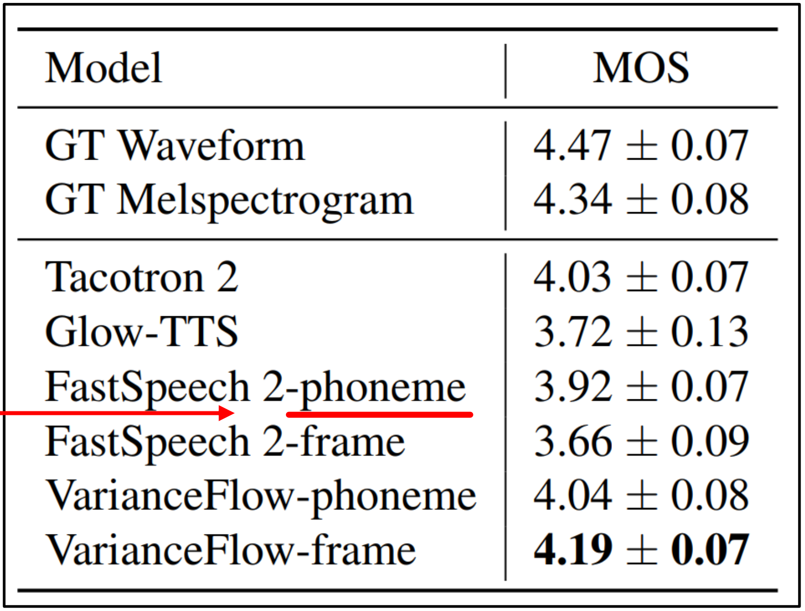


1. Sampling the latent variance representations from the prior
2. Sending the latent representations to the raw variance space using the inverse transforms of NF
3. Adjusting the values in the raw variance space and bringing them back to the latent space using NF modules

Experiments

Experiments

- We measured MOS using Mturk to compare speech quality.



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	$\mathbf{4.19} \pm \mathbf{0.07}$

feeding phoneme-averaged pitch

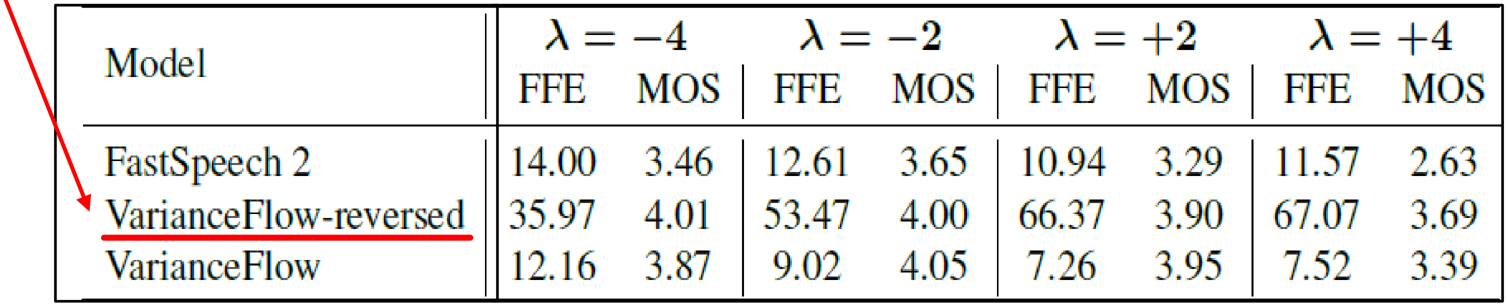
- Our VaraineFlow outperformed the other AR and non-AR TTS models.
- Unlike FastSpeech 2, VarianceFlow benefits from using finer variance information, which verifies its advanced variance modeling ability.

Experiments

- We measured f0 frame error rate (FFE) between input pitch and pitch extracted from generated melspectrogram.

feeding variance information directly

$$\boxed{f_\lambda = 2^{\frac{\lambda}{12}} \times f_0}$$



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

- Our model achieved better pitch controllability with better speech quality.
- The lower MOS in $\lambda = \pm 4$ is because VarianceFlow-reversed does not follow the pitch-shift accurately.

Conclusion

- In this paper. we proposed a novel non-AR TTS model, VarianceFlow.

- By feeding variance information through a NF module, we improved the variance modeling performance and we disentangled the conditioning representations and latent representations

- As a result, we could improve the speech quality and variance controllability of VarianceFlow

[DEMO PAGE]