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.
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 I

x N
Autoregressive TTS

- Autoregressive Text-to-Speech model

\[ p(x_t|x_{<t}, Y) \]

- ‘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

• 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.

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

• 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.

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.

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

- 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
• 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Waveform</td>
<td>4.47 ± 0.07</td>
</tr>
<tr>
<td>GT Melspectrogram</td>
<td>4.34 ± 0.08</td>
</tr>
<tr>
<td>Tacotron 2</td>
<td>4.03 ± 0.07</td>
</tr>
<tr>
<td>Glow-TTS</td>
<td>3.72 ± 0.13</td>
</tr>
<tr>
<td>FastSpeech 2-phoneme</td>
<td>3.92 ± 0.07</td>
</tr>
<tr>
<td>FastSpeech 2-frame</td>
<td>3.66 ± 0.09</td>
</tr>
<tr>
<td>VarianceFlow-phoneme</td>
<td>4.04 ± 0.08</td>
</tr>
<tr>
<td>VarianceFlow-frame</td>
<td><strong>4.19 ± 0.07</strong></td>
</tr>
</tbody>
</table>

• Our VarainceFlow 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.

- 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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda = -4$</th>
<th>$\lambda = -2$</th>
<th>$\lambda = +2$</th>
<th>$\lambda = +4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FFE</td>
<td>MOS</td>
<td>FFE</td>
<td>MOS</td>
</tr>
<tr>
<td>FastSpeech 2</td>
<td>14.00</td>
<td>3.46</td>
<td>12.61</td>
<td>3.65</td>
</tr>
<tr>
<td>VarianceFlow-reversed</td>
<td>35.97</td>
<td>4.01</td>
<td>53.47</td>
<td>4.00</td>
</tr>
<tr>
<td>VarianceFlow</td>
<td>12.16</td>
<td>3.87</td>
<td>9.02</td>
<td>4.05</td>
</tr>
</tbody>
</table>
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.