EfficientTTS: An Efficient and High-Quality Text-to-Speech Architecture

Chenfeng Miao, Shuang Liang, Zhengchen Liu, MinChuan Chen, Jun Ma, Shaojun Wang, Jing Xiao

Ping An Technology

miao_chenfeng@126.com

June 20, 2021

Neural Text-to-Speech Models

Acoustic models (Text-to-Melspectrogram)

- **Autoregressive models**. Tacotron, TransformerTTS, Flowtron, etc.
- **Non-Autoregressive models**. ParaNet, FastSpeech, FastSpeech 2, GlowTTS, etc.

Vocoder (Melspectrogram-to-Waveform)

- **Autoregressive models**. Wavenet, WaveRNN, LPCNet, etc.
- **Non-Autoregressive models**. Parallel-Wavenet, WaveGlow, MelGAN, HiFi-GAN, etc.

End-to-End models (Text-to-Waveform)

- **Autoregressive models**. WaveTacotron
- **Non-Autoregressive models**. Fastspeech 2s, EATS

Proposed approach for Monotonic Alignment Modeling

Index Mapping Vector ($\boldsymbol{\pi}$) is defined as sum of index vector $\boldsymbol{p} = [0, 1, \cdots, T_1 - 1]$, weighted by alignment matrix $\boldsymbol{\alpha}$:

$$\pi_j = \sum_{i=0}^{T_1 - 1} \alpha_{i,j} * p_i$$

Alignment matrix α should follow strict criteria including **Monotonicity**, **Continuity** and **Completeness**.To meet all the criteria, following constraints are true:

$$0 \leq \Delta\pi_i \leq 1$$

$$\pi_0 = 0$$

$$\pi_{T_2 - 1} = T_1 - 1$$

Proposed approach for Monotonic Alignment Modeling

Soft Monotonic Alignment.

$$
\begin{aligned}
\mathcal{L}_{\mathrm{SMA}} &= \lambda_0 \| |\Delta \boldsymbol{\pi}| - \Delta \boldsymbol{\pi} \|_1 \\
&+ \lambda_1 \| |\Delta \boldsymbol{\pi} - 1| + (\Delta \boldsymbol{\pi} - 1) \|_1 \\
&+ \lambda_2 \left(\frac{\pi_0}{T_1 - 1} \right)^2 \\
&+ \lambda_3 \left(\frac{\pi_{T_2 - 1}}{T_1 - 1} - 1 \right)^2
\end{aligned}
$$

Hard Monotonic Alignment.

1.
$$\Delta \pi'_j = \pi'_j - \pi'_{j-1}$$
$$\Delta \pi_j = \mathrm{ReLU}(\Delta \pi'_\mathrm{j})$$
$$\pi_j = \sum_{m=0}^{j} \Delta \pi_m$$
2. $\pi^*_j = \pi_j * \frac{T_1 - 1}{\max(\boldsymbol{\pi})} = \pi_j * \frac{T_1 - 1}{\pi_{T_2 - 1}}$
3. $\alpha'_{i,j} = \frac{\exp\left(-\sigma^{-2}(p_i - \pi^*_j)^2\right)}{\sum_{m=0}^{T_1 - 1} \exp\left(-\sigma^{-2}(p_m - \pi^*_j)^2\right)}$

EfficientTTS Architecture

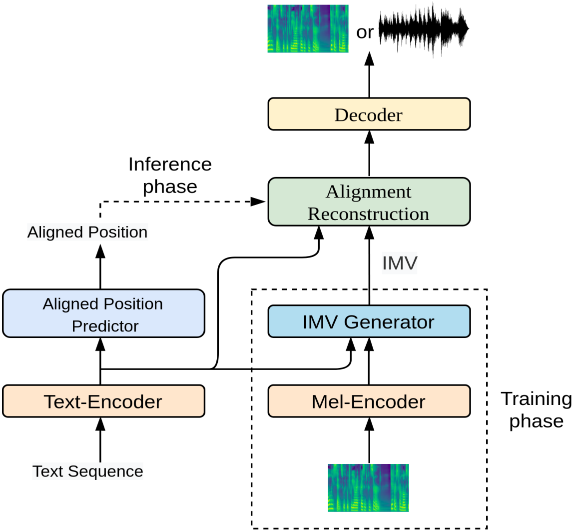

Figure: Overall model architecture.

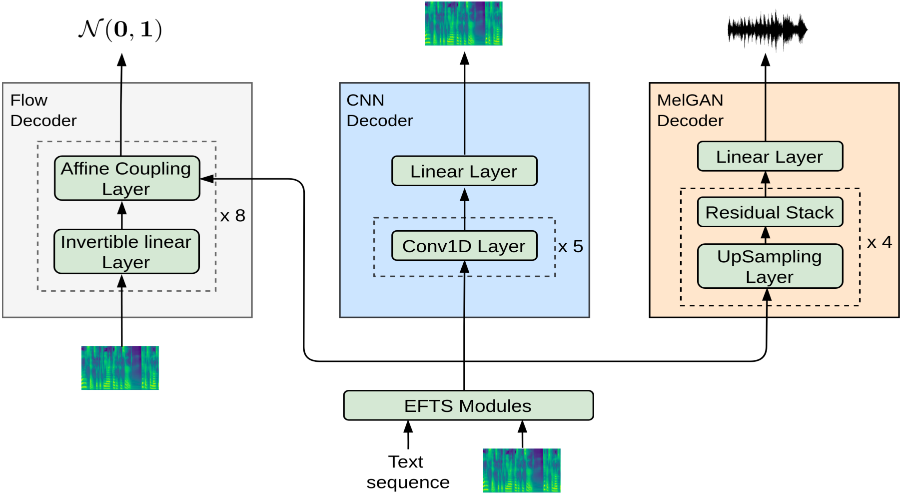

Figure: From left to right are EFTS-Flow, EFTS-CNN and EFTS-Wav respectively.

Experimental results

Table: Quantitative results of training time and inference latency.

Model family	Training Time(h)	Training Speedup	Inference Time text-to-mel(ms)	Inference Speedup text-to-mel	Inference Time text-to-wav(ms)	Inference Speedup text-to-wav
Tacotron 2	54	-	780	-	824	-
Glow-TTS	120	0.45×	42	18.6×	86	9.6×
EFTS-CNN	6	**9**×	**8**	**97.5**×	52	15.8×
EFTS-Flow	32	1.7×	21	37.1×	65	12.7×
EFTS-Wav	-	-	-	-	**18**	**45.8**×

Table: MOS on DataBaker.

Method	MOS
Ground Truth	4.64 ± 0.07
Ground Truth (Mel+HiFi-GAN)	4.58 ± 0.13
Tacotron 2 (Mel+HiFi-GAN)	4.20 ± 0.11
Glow-TTS (Mel+HiFi-GAN)	3.97 ± 0.21
EFTS-CNN (Mel+HiFi-GAN)	$\mathbf{4.41 \pm 0.13}$
EFTS-Flow (Mel+HiFi-GAN)	$\mathbf{4.35 \pm 0.17}$
EFTS-Wav	$\mathbf{4.40 \pm 0.21}$

Table: MOS on LJ-Speech.

Method	MOS
Ground Truth	4.75 ± 0.12
Ground Truth (Mel+HiFi-GAN)	4.51 ± 0.13
Tacotron 2 (Mel+HiFi-GAN)	4.08 ± 0.13
Glow-TTS (Mel+HiFi-GAN)	4.13 ± 0.18
EFTS-CNN (Mel+HiFi-GAN)	$\mathbf{4.27 \pm 0.14}$

The End