TTS approaches comparison

Text-to-speech synthesis

$$W_k$$

$$Y_k = \mathbf{y}_{k,1} \cdots \mathbf{y}_{k,n} \cdots \mathbf{y}_{k,N}$$

$$Z = \mathbf{z}_1 \cdots \mathbf{z}_m \cdots \mathbf{z}_M$$

$$S$$

Q1: What is the shared latent symbol set $\{a^d\}_{d=1}^{D}$?

Q2: How to map Z to latent symbol sequence S?

Q3: How to map W_k to latent symbol sequence Y_k?

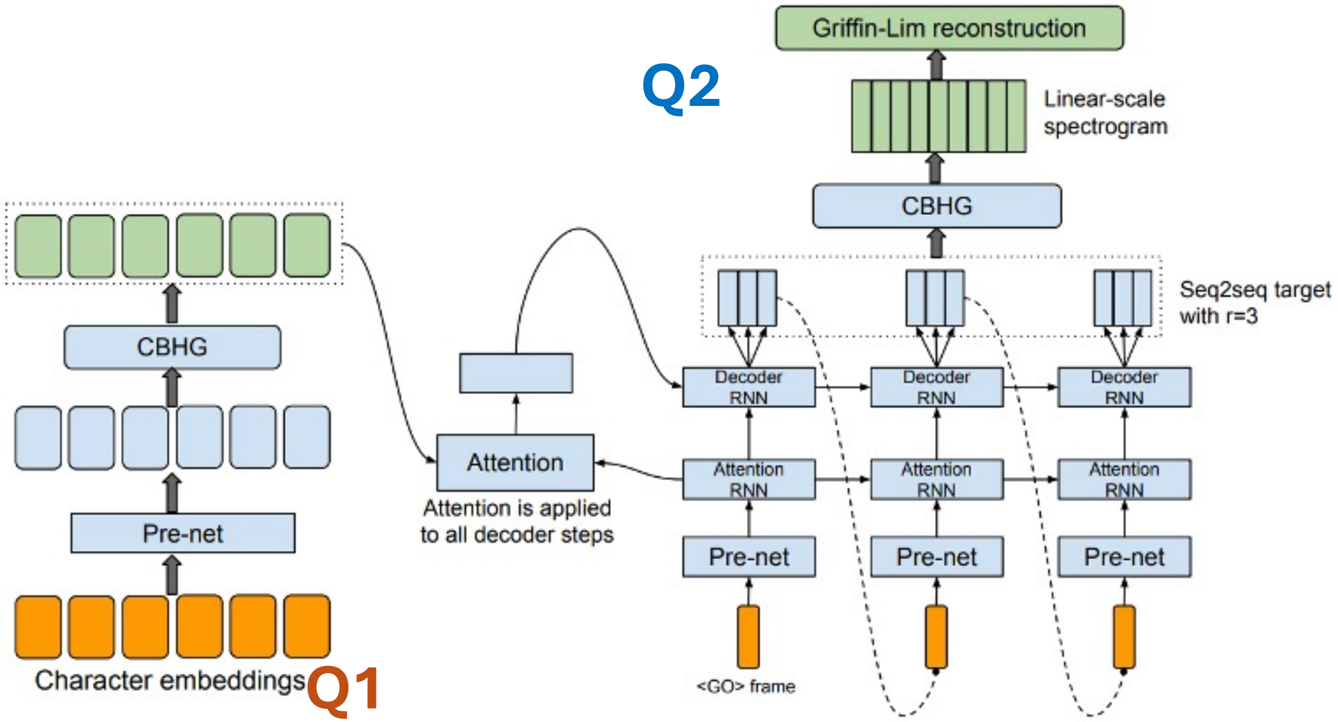
Q4: Map Y_k to Z (explicitly/implicitly integrates a seq. matching process).

Comparison between TTS approaches

	Q1	Q3	Q4	Q2
Unit sel. concatenative synthesis	Diphones	Text-to-diphone target generation	Viterbi algo. for unit selection and concatenation	Signal processing, e.g., PSOLA
HMM-based synthesis	Clustered context-dependent phones	Text-to-clustered CD phone state seq. generation	Maximum likelihood generation of (source-system) vocoder parameters	Vocoding
Neural TTS	Neural embeddings	Text to seq. of linguistic embeddings	Linguistic embedding seq. to acoustic representation seq.	Neural vocoding

Neural TTS: two stage approach

Tacotron



Q3

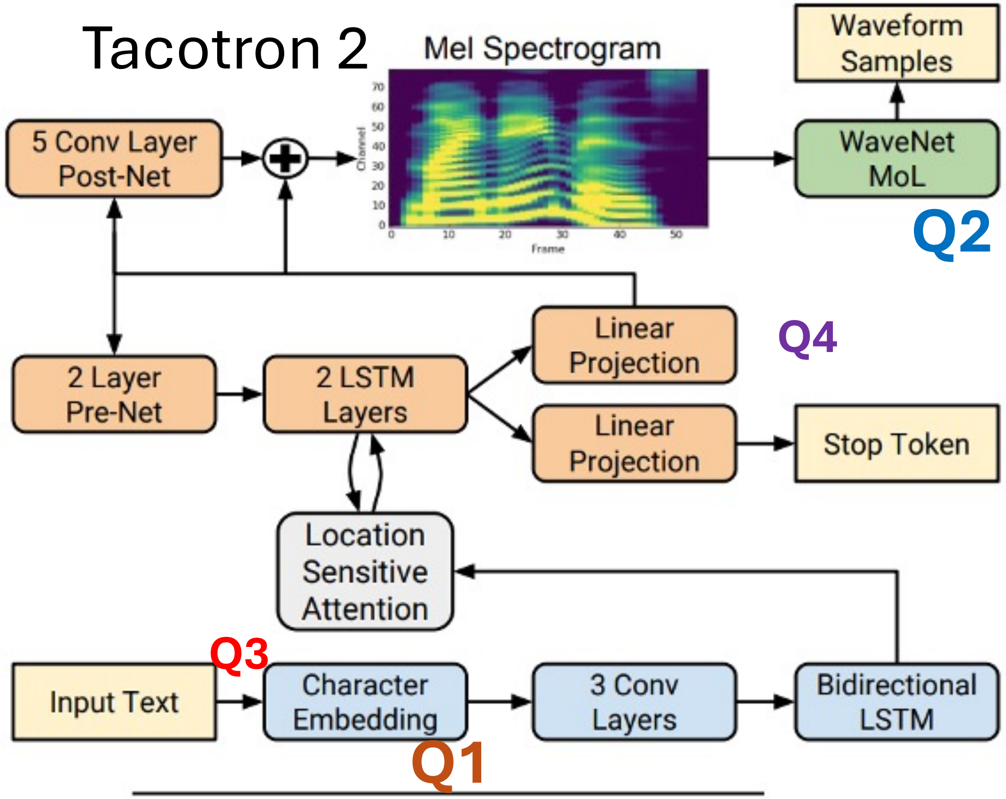
Text-to-Character embeddings

Q4

	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

Wang et al. Tacotron: Towards End-to-End Speech Synthesis, Proc. Interspeech, 2017

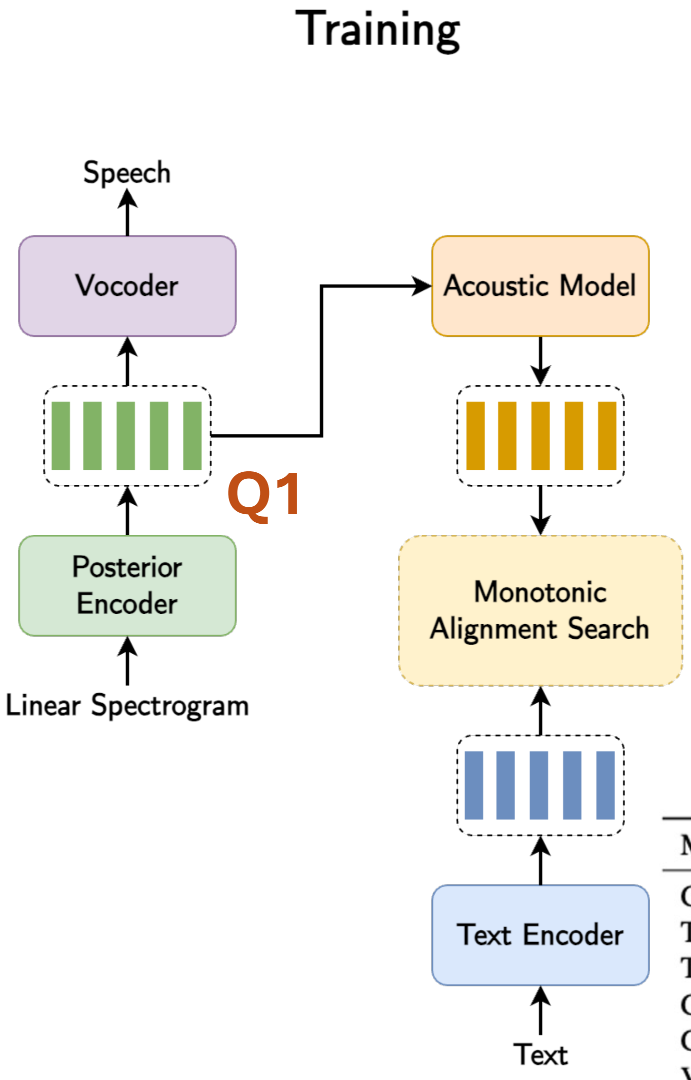
Tacotron 2

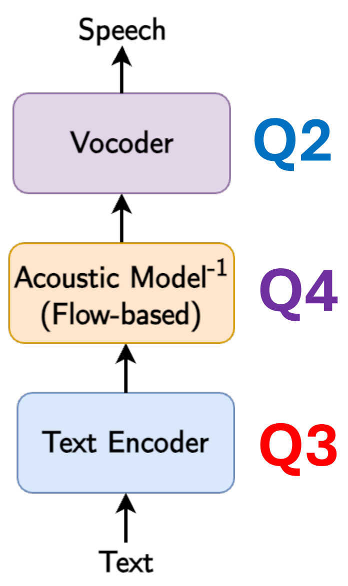


System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Shen et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, Proc. ICASSP, 2018

Neural TTS: end-to-end approach



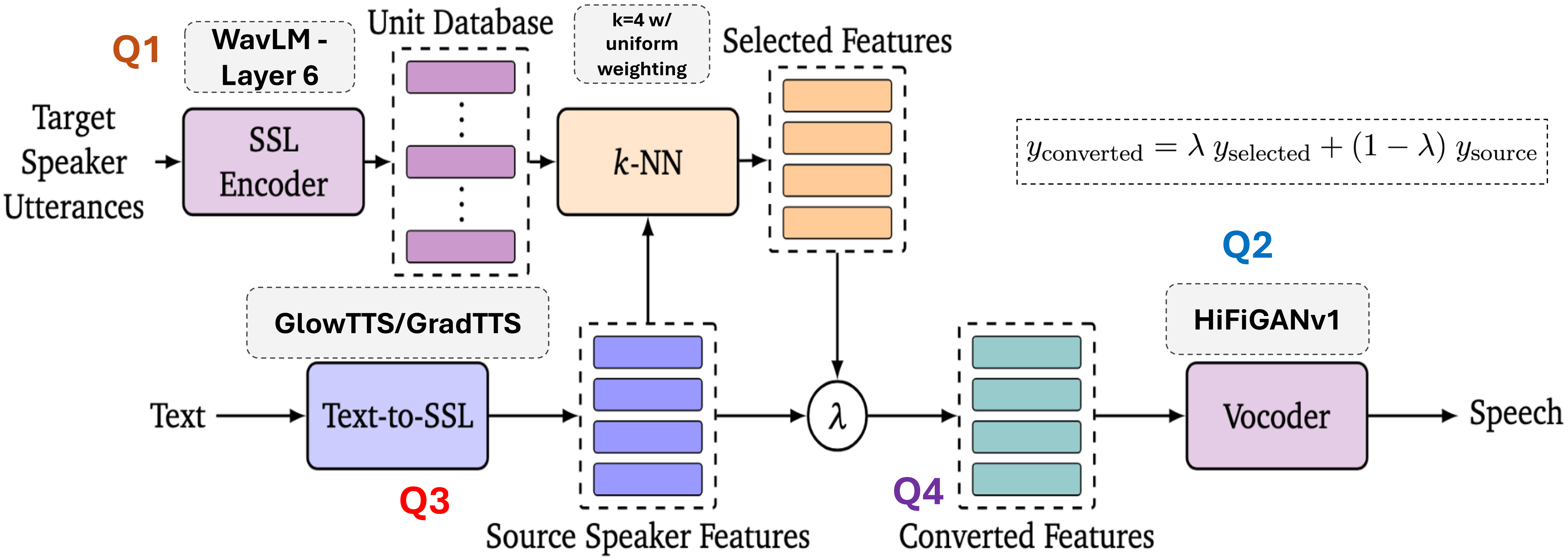


Model	MOS (CI)
Ground Truth	4.46 ($\pm$0.06)
Tacotron 2 + HiFi-GAN	3.77 ($\pm$0.08)
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 ($\pm$0.07)
Glow-TTS + HiFi-GAN	4.14 ($\pm$0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 ($\pm$0.07)
VITS (DDP)	4.39 ($\pm$0.06)
VITS	**4.43 ($\pm$0.06)**

- **Two-stage**
  - <u>Pros:</u> interpretable intermediate representation, modular way, vocoder can be trained on untranscribed speech data
  - <u>Cons:</u> suffer from error propagation, handcrafted feature limitations
- **End-to-end**
  - <u>Pros:</u> simplified joint training, less error propagation, achieve higher naturalness
  - <u>Cons:</u> reduced flexibility, less interpretable, can suffer from oversmoothing and mispronunciation

Kim, Kong and Son,"Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech", in Proc. of ICML, 2021

kNN-based multi-speaker TTS (1)



λ = 0 λ = 0.25 λ = 0.5 λ = 0.75 λ = 1 Ground truth

K. El Hajal, A. Kulkarni, E. Hermann and M. Magimai-Doss, "kNN Retrieval for Simple and Effective Zero-Shot Multi-speaker Text-to-Speech", in Proc. of NAACL, 2025