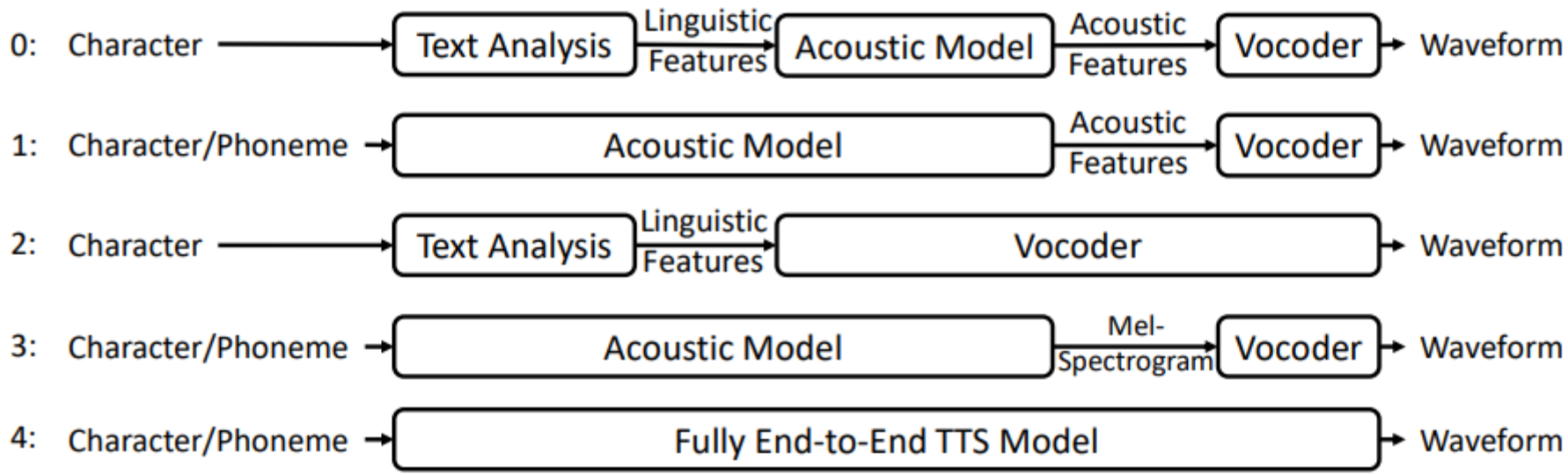
Современный синтез речи

(2023)



Stage	Models
0	SPSS [416, 356, 415, 425, 357]
1	ARST [375]
2	WaveNet [254], DeepVoice 1/2 [8, 87], Par. WaveNet [255], WaveRNN [150], HiFi-GAN [23]
3	DeepVoice 3 [270], Tacotron 2 [303], FastSpeech 1/2 [290, 292], WaveGlow [279], FloWaveNet [163]
4	Char2Wav [315], ClariNet [269], FastSpeech 2s [292], EATS [69], Wave-Tacotron [385], VITS [160]

Figure 4: The progressively end-to-end process for TTS models.

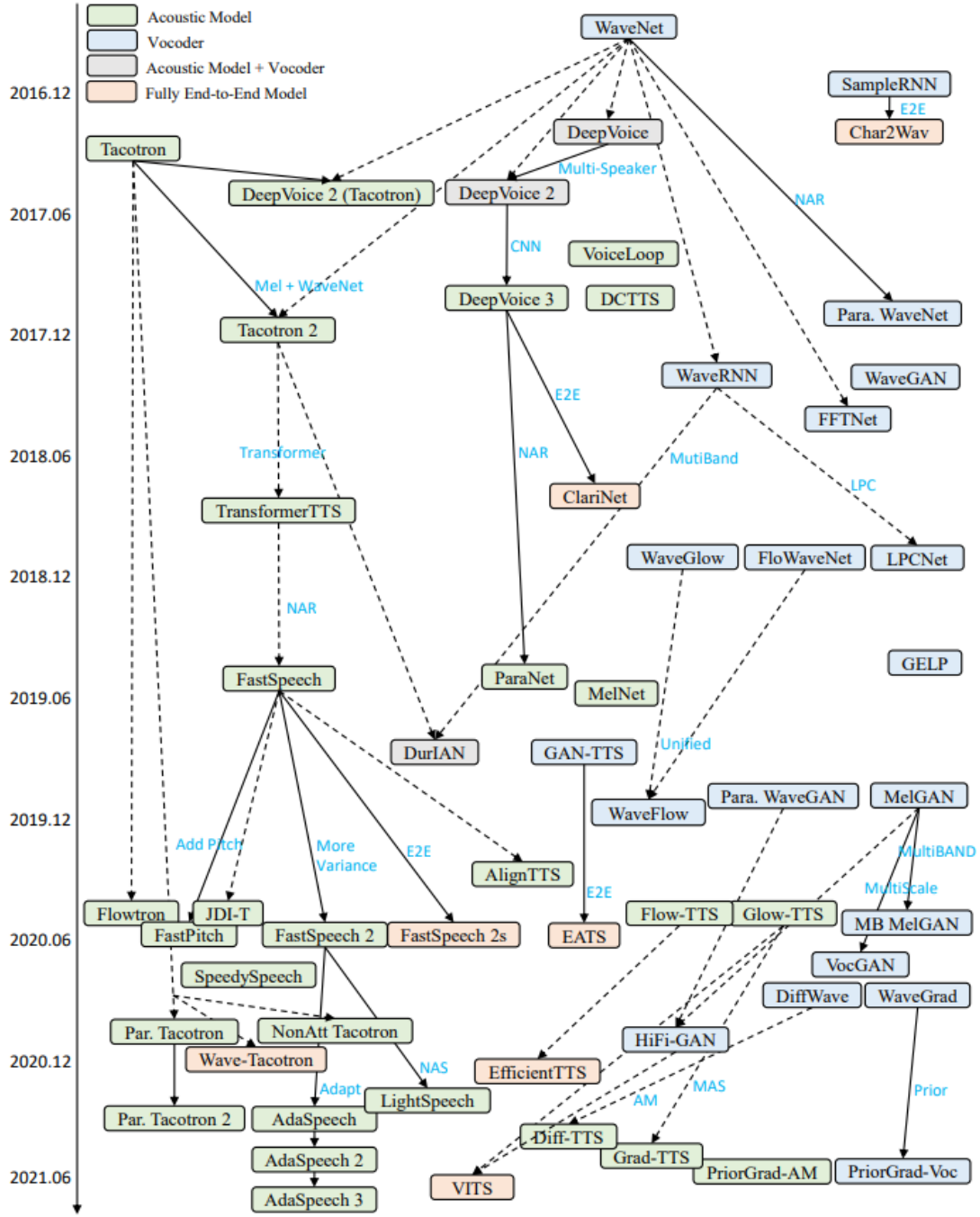
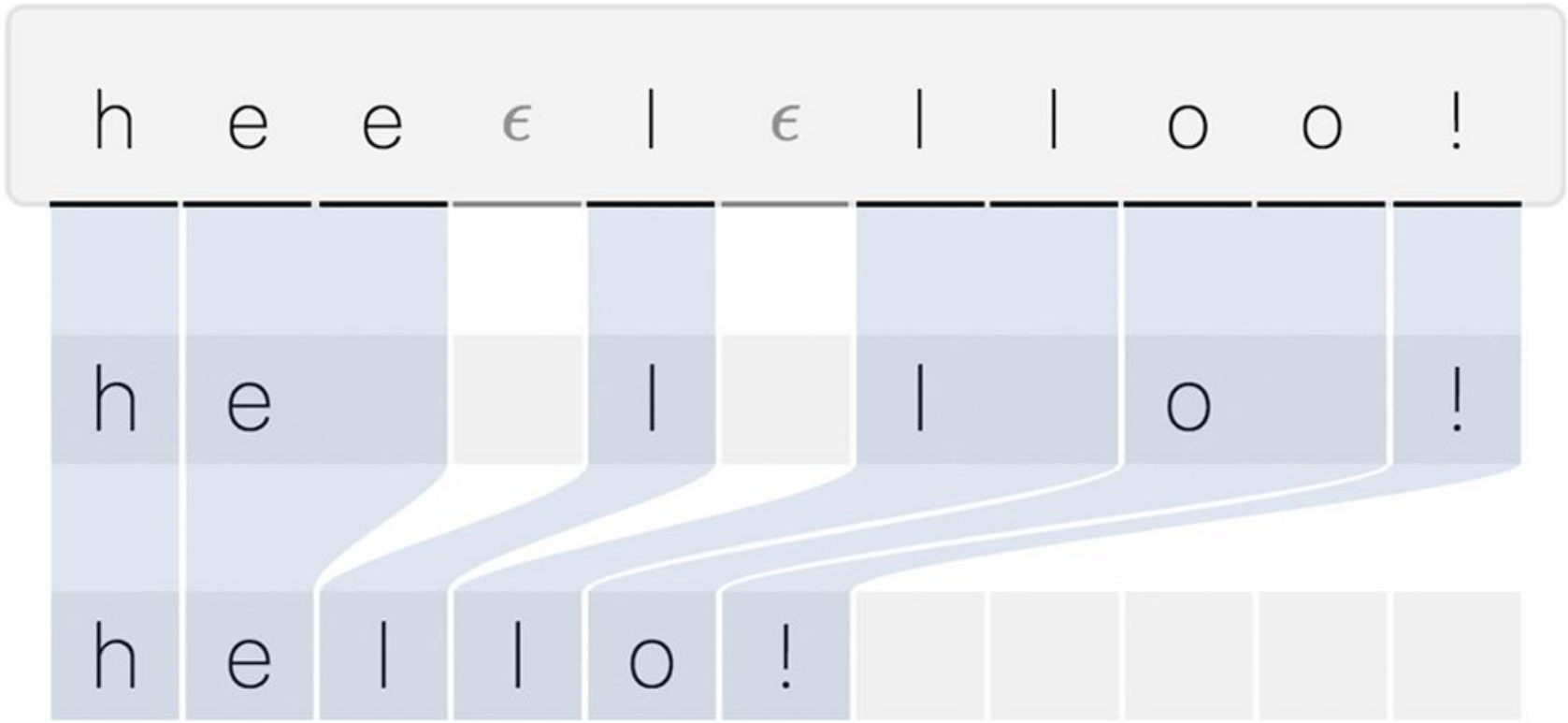
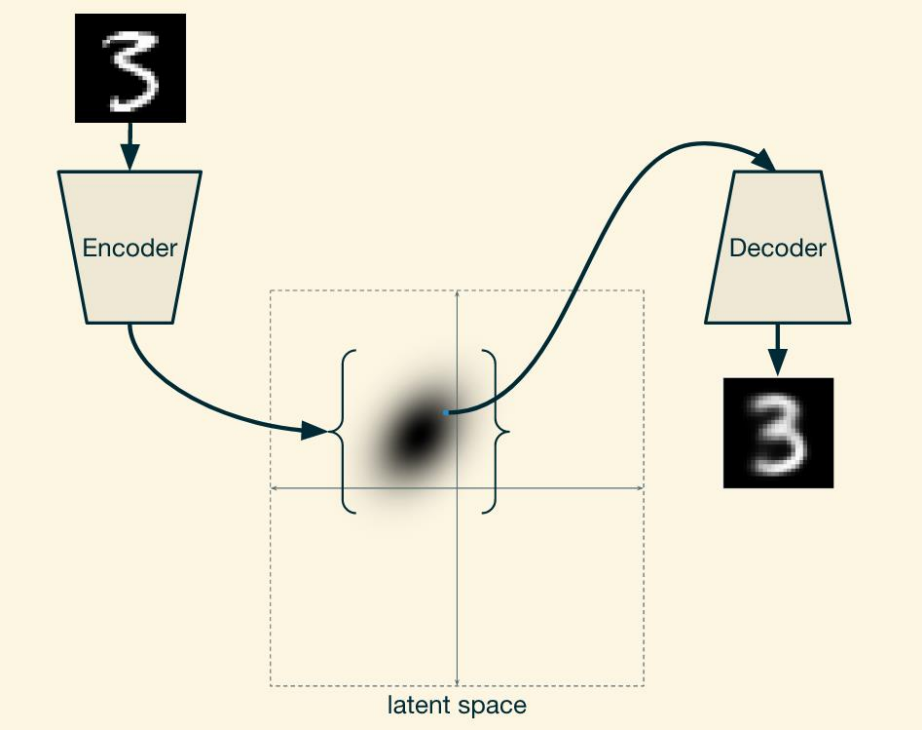


Figure 6: The evolution of neural TTS models.

Alignment

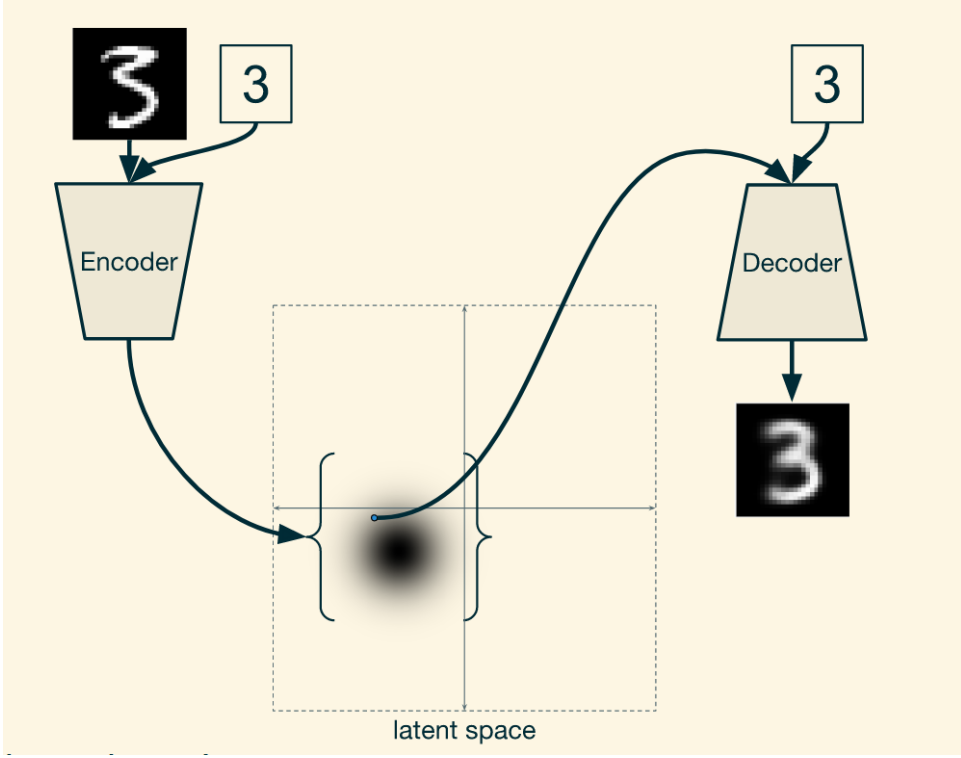


VAE/CVAE

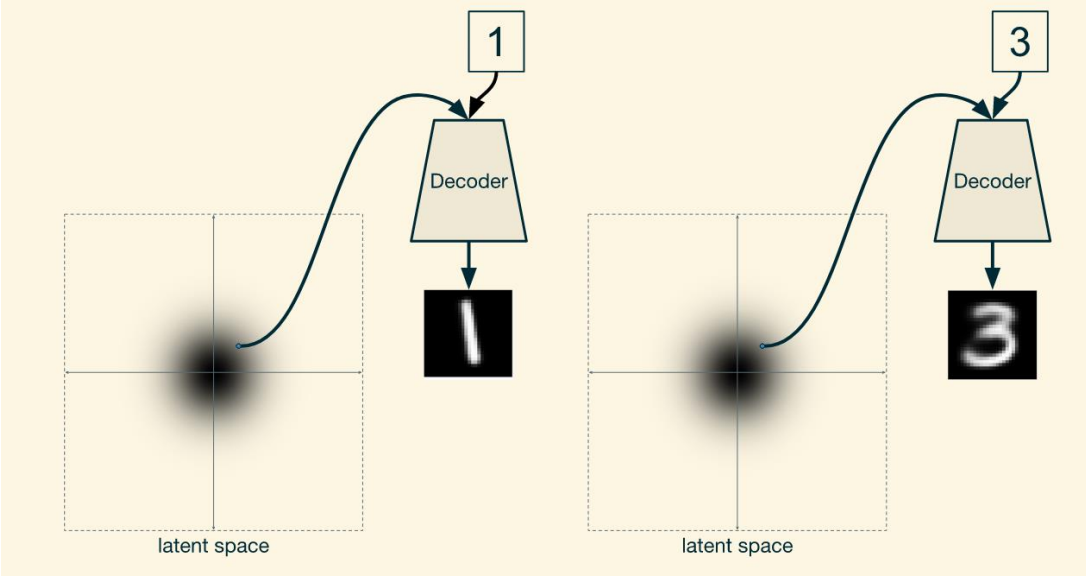


VAE

CVAE
training



CVAE
inference

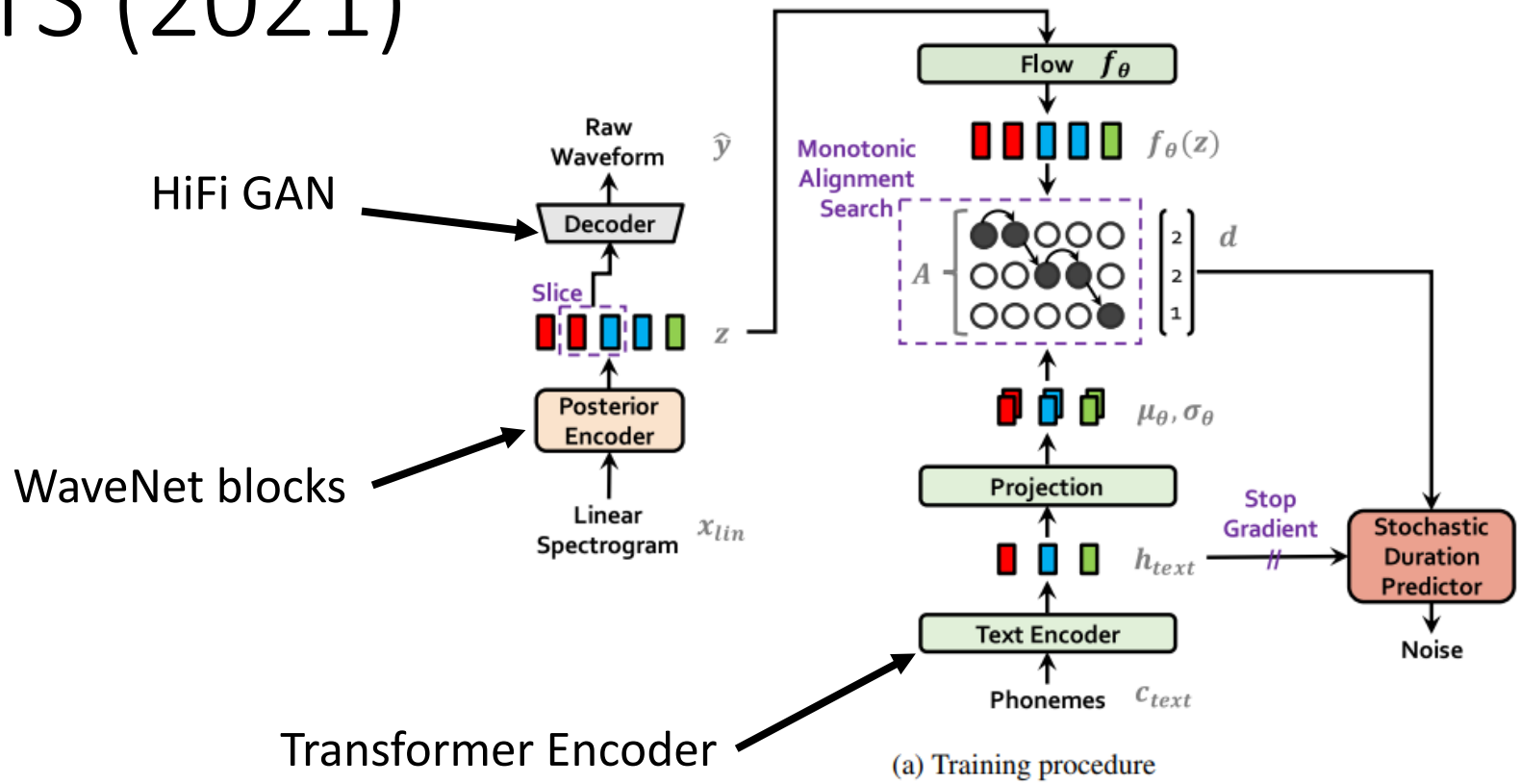


VITS (2021)

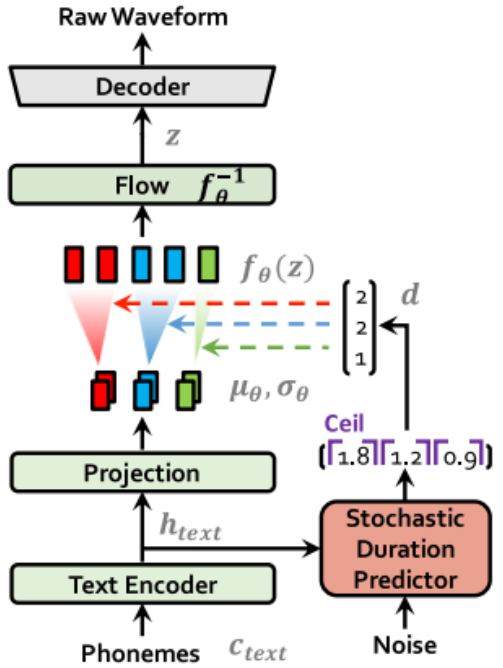
HiFi GAN

WaveNet blocks

Transformer Encoder



(a) Training procedure



(b) Inference procedure

phoneme → waveform

VITS (2021)

Table 1. Comparison of evaluated MOS with 95% confidence intervals on the LJ Speech dataset.

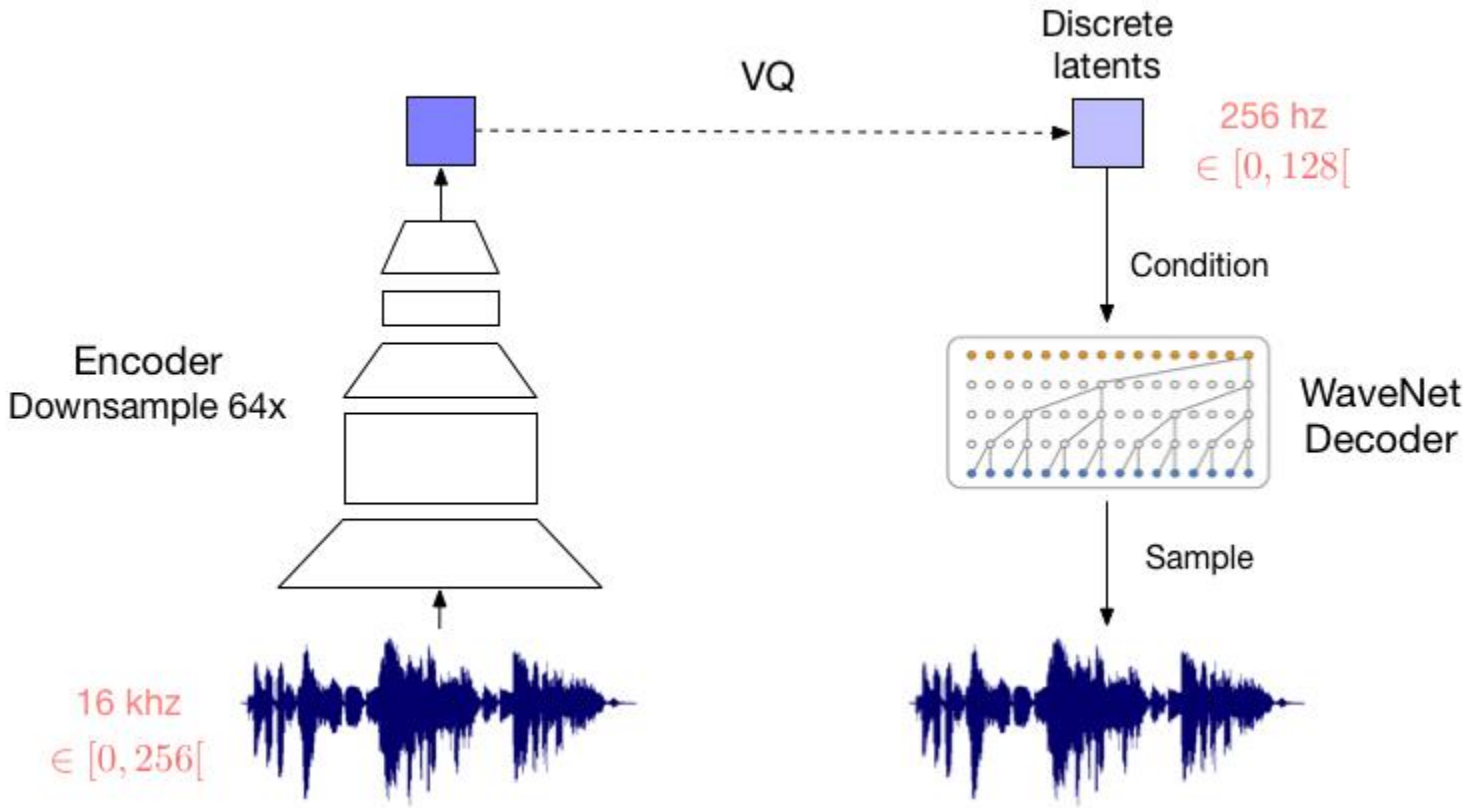
Model	MOS (CI)
Ground Truth	4.46 (±0.06)
Tacotron 2 + HiFi-GAN	3.77 (±0.08)
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 (±0.07)
Glow-TTS + HiFi-GAN	4.14 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 (±0.07)
VITS (DDP)	4.39 (±0.06)
VITS	**4.43 (±0.06)**

Table 3. Comparison of evaluated MOS with 95% confidence intervals on the VCTK dataset.

Model	MOS (CI)
Ground Truth	4.38 (±0.07)
Tacotron 2 + HiFi-GAN	3.14 (±0.09)
Tacotron 2 + HiFi-GAN (Fine-tuned)	3.19 (±0.09)
Glow-TTS + HiFi-GAN	3.76 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	3.82 (±0.07)
VITS	**4.38 (±0.06)**

https://jaywalnut310.github.io/vits-demo/index.html

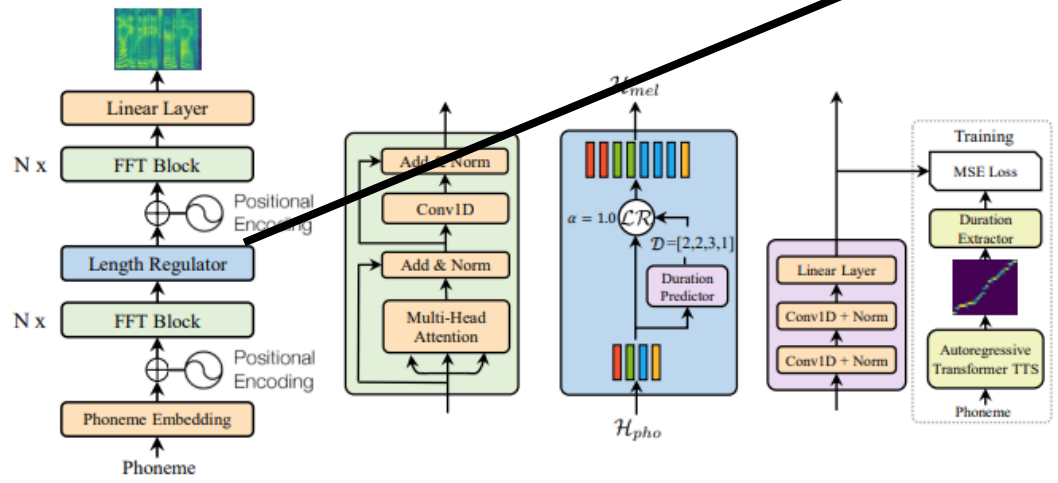
VQ-VAE


VQ
Discrete
latents
256 hz
$\in [0, 128[$
Condition
Encoder
Downsample 64x
WaveNet
Decoder
Sample
16 khz
$\in [0, 256[$

NaturalSpeech (2022)

Phoneme → waveform

VQ-VAE



(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Figure 1: The overall architecture for FastSpeech. (a). The feed-forward Transformer. (b). The feed-forward Transformer block. (c). The length regulator. (d). The duration predictor. MSE loss denotes the loss between predicted and extracted duration, which only exists in the training process.

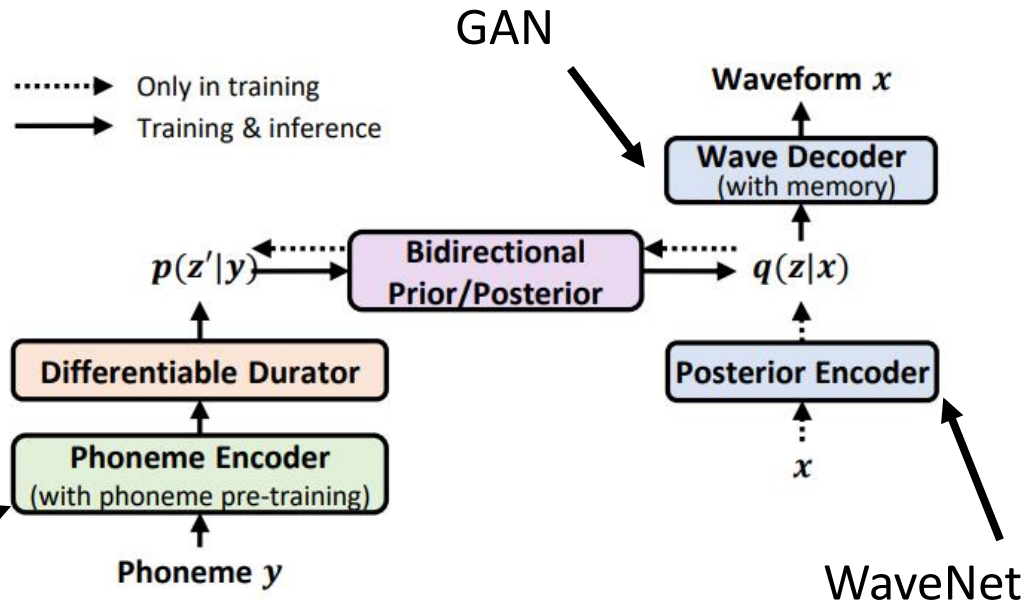
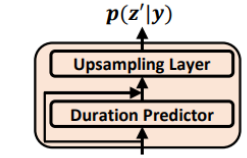
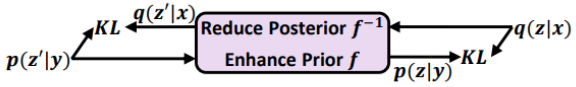


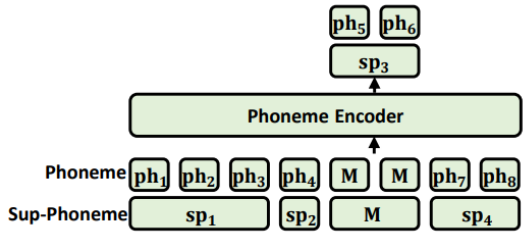
Figure 1: System overview of NaturalSpeech.



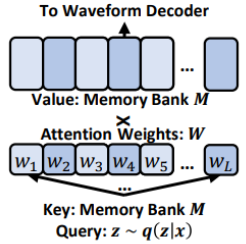
(a) Differentiable durator.



(b) Bidirectional prior/posterior.



(c) Phoneme pre-training.



(d) Memory mechanism in VAE.

Figure 2: The designed modules in NaturalSpeech.

NaturalSpeech (2022)

Table 4: MOS and CMOS comparisons between NaturalSpeech and previous TTS systems.

System	MOS	CMOS
FastSpeech 2 [18] + HiFiGAN [17]	4.32 ± 0.15	-0.33
Glow-TTS [13] + HiFiGAN [17]	4.34 ± 0.13	-0.26
Grad-TTS [14] + HiFiGAN [17]	4.37 ± 0.13	-0.24
VITS [15]	4.43 ± 0.13	-0.20
NaturalSpeech	4.56 ± 0.13	0

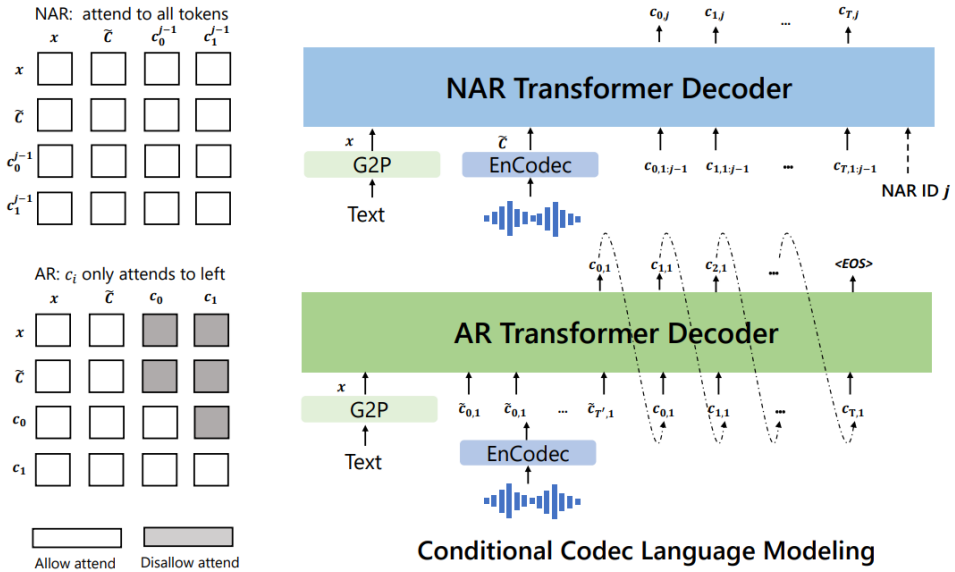
- “which demonstrates no statistically significant difference from human recordings for the first time on this dataset.”

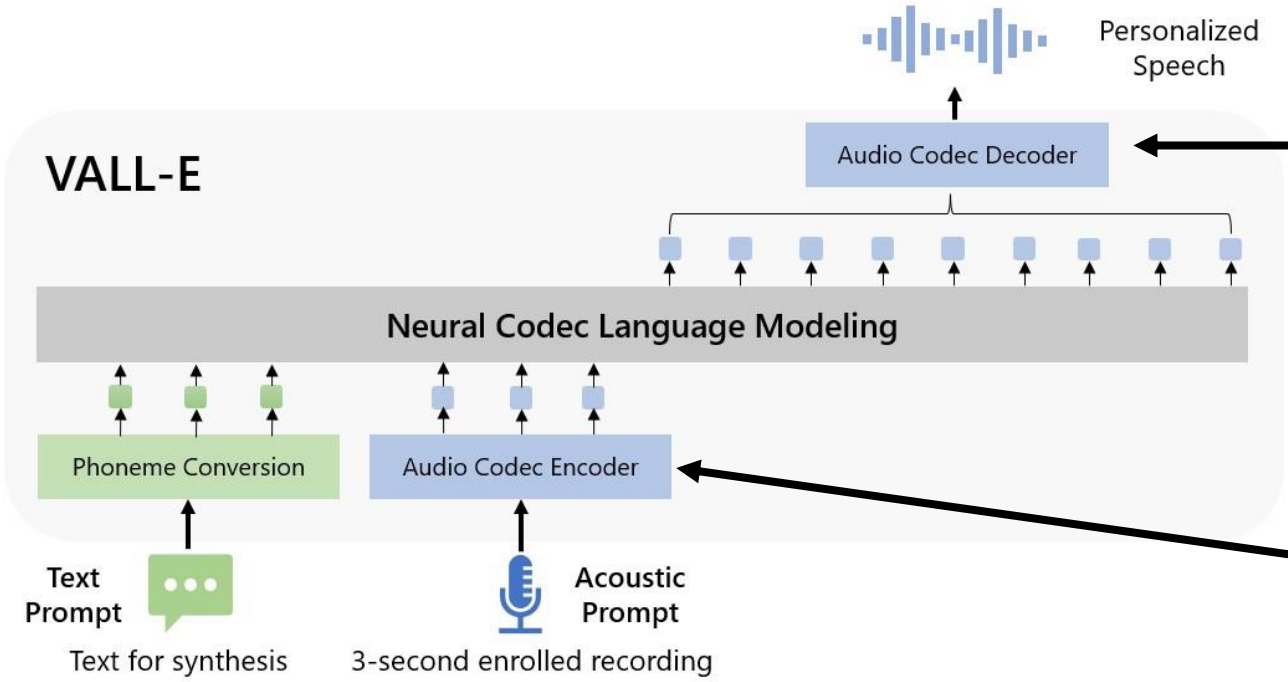
https://speechresearch.github.io/naturalspeech/

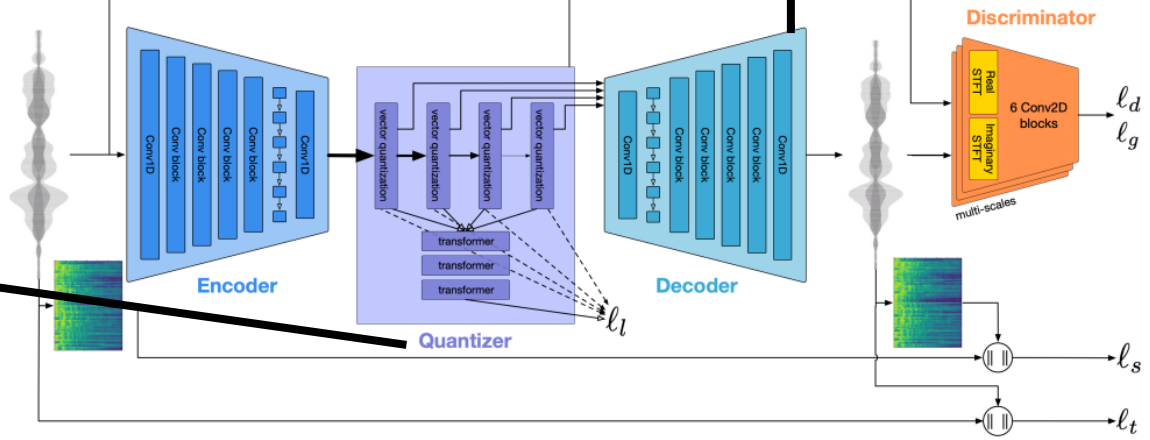
VALL-E (2023)

phoneme → discrete code → waveform

- neural codec language model
- Pre-training 60k hours







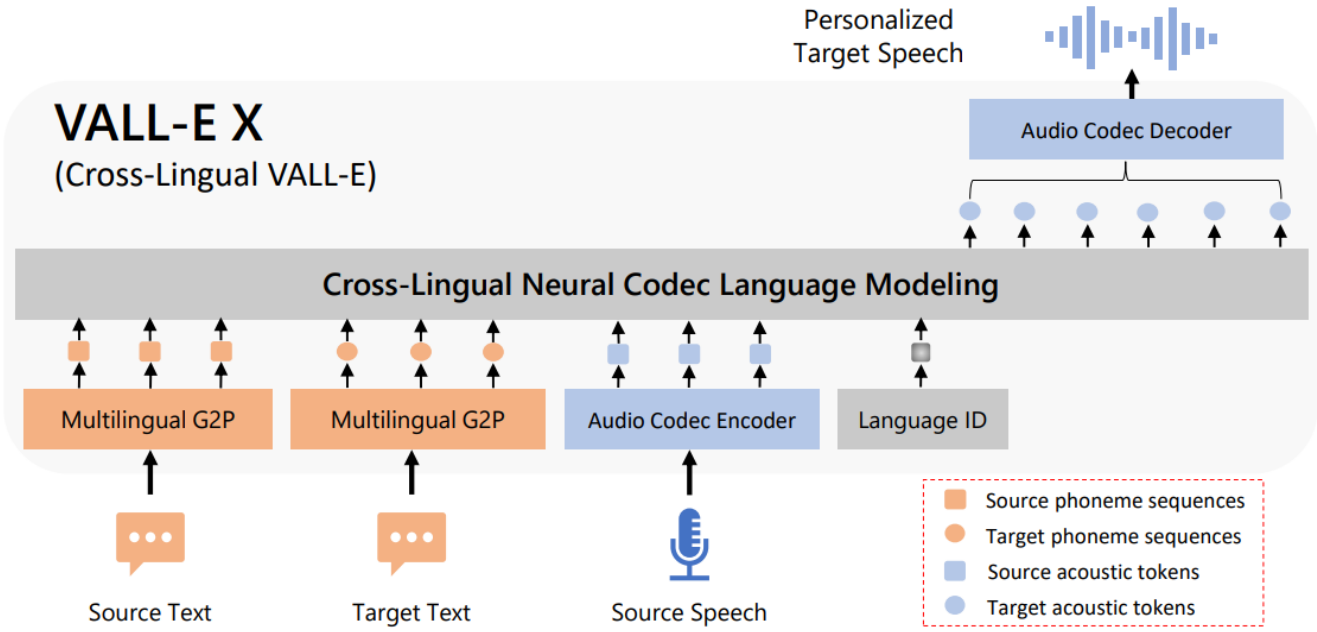
VALL-E (2023)

Table 7: Human evaluation with 60 speakers on VCTK with 3-second enrolled recording for each.

	SMOS	CMOS (v.s. VALL-E)
YourTTS*	$3.70_{\pm 0.09}$	-0.23
VALL-E	$3.81_{\pm 0.09}$	0.00
GroundTruth	$4.29_{\pm 0.09}$	-0.04

https://valle-demo.github.io/

VALL-E X (2023)



- https://vallex-demo.github.io/

Список использованной литературы

- https://arxiv.org/pdf/2106.15561.pdf
- https://arxiv.org/pdf/2106.06103.pdf
- https://arxiv.org/pdf/2205.04421.pdf
- https://arxiv.org/pdf/2301.02111.pdf