Recent advances in TTS: from specific to universal solutions

Evgeny Lagutin, GigaChat TTS team
tg: @elagutin

Shortcomings of outdated models

1. Limited Quality & Naturalness
   - Fair, but often flat or monotonous intonation
   - Insufficient naturalness and presence of artifacts
   - Synthetic sound that fails to mimic human voice nuances
2. Lack of Flexibility & Adaptability
   - Poor zero-shot capabilities
   - No voice / domain / task adaptation
   - Minimal user control over speaking style, emotion, or prosody
   - Absence of transfer learning, no effective knowledge sharing between different voices or tasks
3. Architectural Limitations
   - Lacks pretrained audio representations for fast fine-tuning
   - Models are trained from scratch for specific tasks, limiting generalization
   - Separate models for text processing, acoustic modeling, and vocoding lead to error accumulation

Towards human-level speech quality

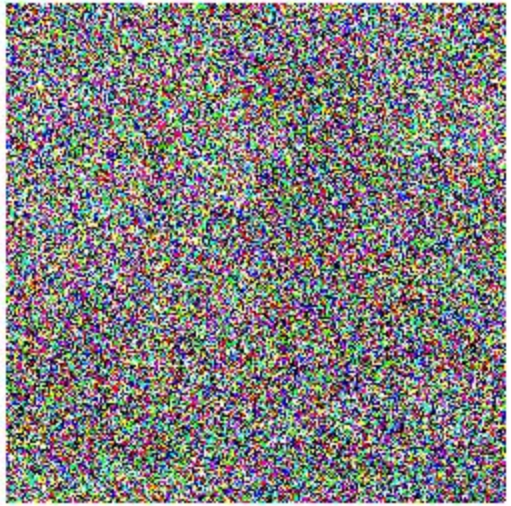
1. **Data Revolution: Scale and Diversity**
   - Availability of large-scale, open-source speech corpora (e.g., LibriLight, MLS, Emilia)
   - Unsupervised/Self-Supervised Learning (HuBERT, WavLM, EnCodec)
   - Training on diverse, real-world data improves robustness and naturalness

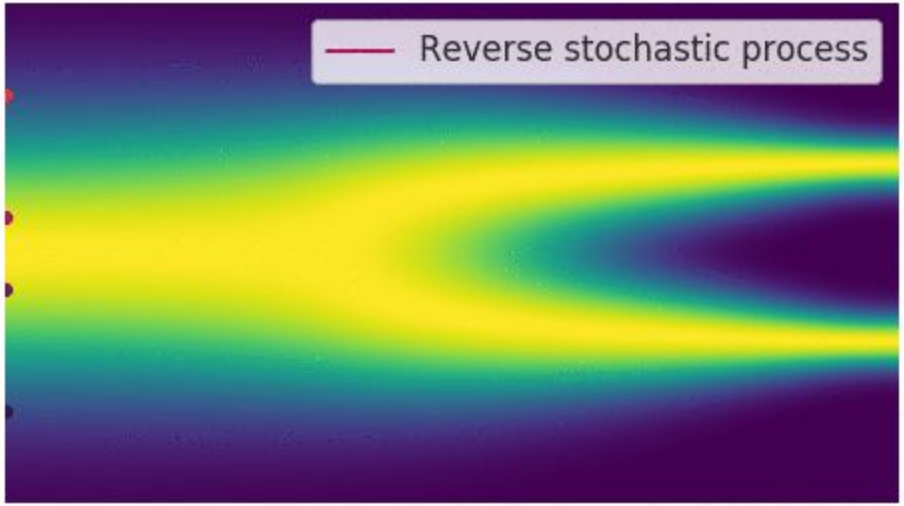
2. **Computational Power: Scaling Model and Inference**
   - Multi-GPU/TPU clusters enable training of billion-parameter models
   - GPU servers and frameworks make real-time synthesis feasible for mass applications
     Efficient Architectures designed for faster inference without sacrificing quality

3. **Architectural Breakthroughs**
   - Transformers excel at capturing long-range dependencies in both text and audio sequences
   - In-Context Learning, massive pretraining and prompting enable generalization on new tasks and domains
   - Denoising Diffusion models can model high-fidelity continuous signals

4. **Unified Representations and Training**
   - Audio Codecs simplify the generation task
   - Emergence of foundation models capable of multiple tasks (TTS, ASR) within a single architecture

Denoising Diffusion: Quick Overview





Denoising Diffusion: Quick Overview

Forward process:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I})$$

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Then:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad \text{and} \quad q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

Model of the reverse process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

where $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$. Although true reverse process $q(x_{t-1}|x_t)$ is intractable, still

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t \mathbf{I}),$$

where

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \quad \text{and} \quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

Denoising Diffusion: Quick Overview

The ojective is a variational lower bound of log-likelihood of the data:

$$L_{VLB} = \mathbb{E}_q \left[D_{KL}(q(x_T|x_0) \mid p(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \mid p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \right]$$

Both distributions inside KL are Gaussians with the same variance, thus:

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} |\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)|^2 \right] + C$$

The form of $\tilde{\mu}_t$ can be expressed as

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right)$$

Therefore, we can parameterize our model to predict this noise. The final, simple **DDPM training objective**:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]$$

Denoising Diffusion: Quick Overview

Training:

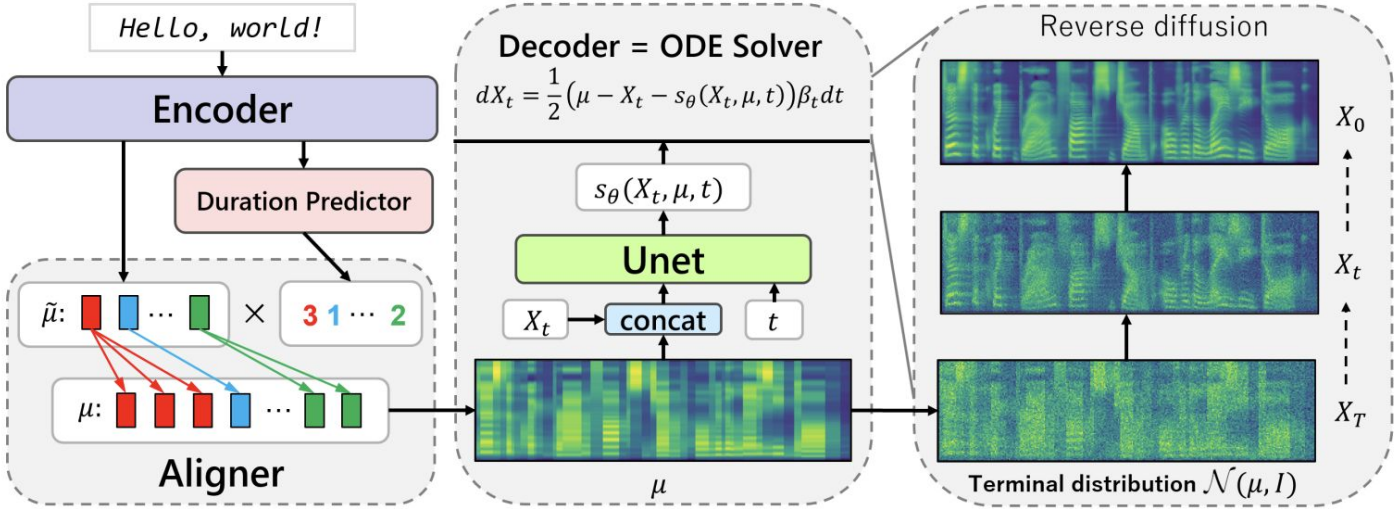
1. Sample a clean data point x_0.
2. Sample a timestep $t \sim \mathrm{Uniform}(1, T)$.
3. Sample random noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.
4. Take a gradient descent step on $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2$.

Sampling (Generation):

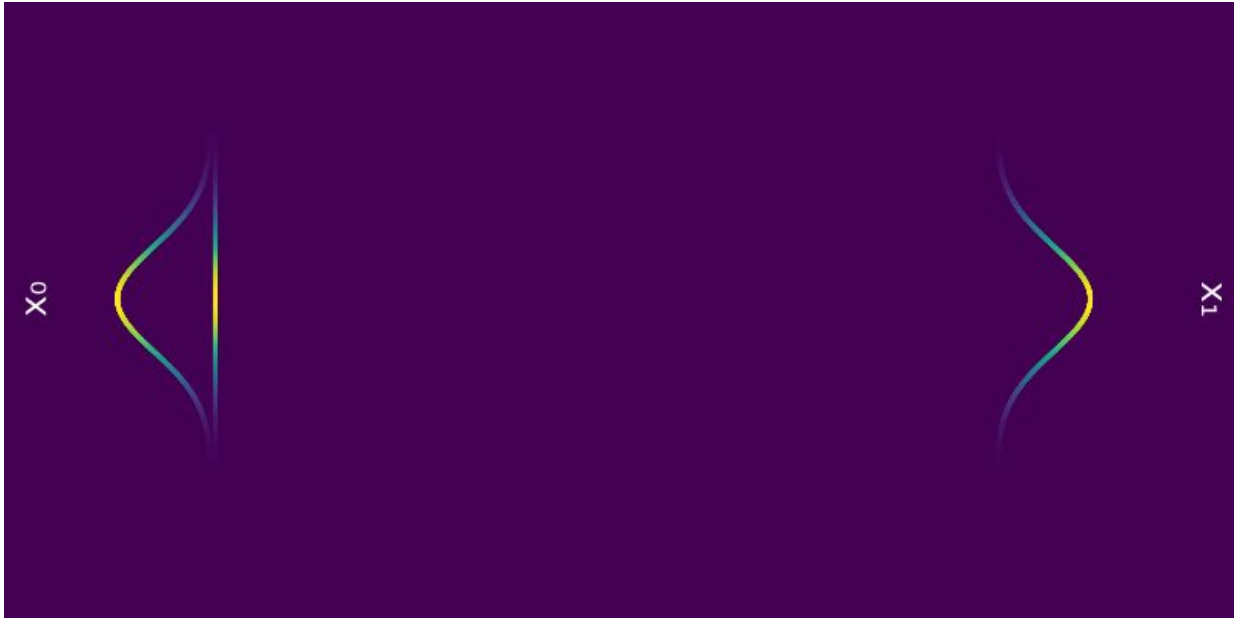
1. Sample $x_T \sim \mathcal{N}(0, \mathbf{I})$.
2. For $t = T$ down to 1:
   - Predict the noise: $\epsilon_\theta = \epsilon_\theta(x_t, t)$
   - Get the mean: $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\right)$
   - Sample the previous step: $x_{t-1} \sim \mathcal{N}(\mu_\theta(x_t, t), \sigma_t^2 \mathbf{I})$, where σ_t^2 is set to $\tilde{\beta}_t$.
3. Return x_0.

Grad-TTS [Huawei Noah Ark, ‘21]

- First diffusion-based TTS method
- Duration predictor adapted from Glow-TTS
- Prior loss added to quickly learn to align input symbols with output frames
- High quality by the cost of slow inference



Conditional Flow Matching: Quick Overview



Conditional Flow Matching: Quick Overview

Let $p_0 \equiv \mathcal{N}(0, \mathbf{I})$, and p_1 be the data distribution. For a flow ϕ_t, s.t. $\phi_0(x) = x$, the density of $p_t = (\phi_t)_\sharp p_0$ can be derived via the change of variables formula:

$$p_t(x) = p_0\left(\phi_t^{-1}(x)\right) \det\left[\frac{\partial \phi_t^{-1}}{\partial x}(x)\right]$$

The vector (velocity) field $u_t(x)$ is defined by:

$$\frac{d}{dt}\phi_t(x) = u_t(\phi_t(x)), \quad \phi_0(x) = x$$

Let p_t be a probability path and u_t be the corresponding vector field that generates p_t. The vector field $u_\theta(x, t)$ parameterized by θ can be trained with the Flow Matching objective:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t\sim U(0,1), x\sim p_t(x)} \|u_\theta(x, t) - u_t(x)\|^2$$

This is intractable because we don't know $p_t(x)$ or $u_t(x)$. Still p_t may be constructed as a mixture of distributions $p_t(x \mid x_1)$. For each x_1, we define $p_t(x \mid x_1)$ such that: $p_0(x \mid x_1) = p_0(x)$ and $p_1(x \mid x_1) \approx \delta(x - x_1)$. Let $u_t(x \mid x_1)$ generate this path. The key theorem states that the gradients of

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t\sim U(0,1), x_1\sim p_1, x\sim p_t(x|x1)} \|u_\theta(x, t) - u_t(x \mid x_1)\|^2$$

are equal to the gradients of the intractable $\mathcal{L}_{FM}(\theta)$ up to a constant.

Conditional Flow Matching: Quick Overview

A common and effective choice for flow is an **affine transformation**:

$$x_t = \psi_t(x_0 \mid x_1) = \sigma_t x_0 + \mu_t x_1$$

The conditional vector field $u_t(x \mid x_1)$ is the time derivative of this flow:

$$u_t(x_t \mid x_1) = \frac{d}{dt}\psi_t(x_0 \mid x_1) = \dot{\sigma}_t x_0 + \dot{\mu}_t x_1 = \dot{\sigma}_t \frac{x_t - \mu_t x_1}{\sigma_t} + \dot{\mu}_t x_1 = \frac{\dot{\sigma}_t}{\sigma_t} x_t + \left(\dot{\mu}_t - \frac{\dot{\sigma}_t \mu_t}{\sigma_t}\right) x_1$$

For the quadratic transport cost $c(x_0, x_1) = \frac{1}{2}\|x_0 - x_1\|^2$, the optimal coupling π^* minimizes:

$$\mathbb{E}_{(x_0,x_1)\sim\pi}\left[\frac{1}{2}\|x_0 - x_1\|^2\right]$$

subject to the marginal constraints $\pi(\cdot, \mathcal{X}_1) = p_0$ and $\pi(\mathcal{X}_0, \cdot) = p_1$.

With the linear scheduler ($\sigma_t = 1 - t, \mu_t = t$), the conditional flow and vector field simplify dramatically:

- Flow: $x_t = (1 - t)x_0 + tx_1$
- Conditional Vector Field: $u_t(x_t \mid x_1) = x_1 - x_0$

The **OT-CFM objective** is then:

$$\mathcal{L}_{OT-CFM}(\theta) = \mathbb{E}_{t\sim U(0,1),(x_0,x_1)\sim\pi^*}\|u_\theta(x_t, t) - (x_1 - x_0)\|^2$$

Conditional Flow Matching: Quick Overview

Training:

For each training iteration:

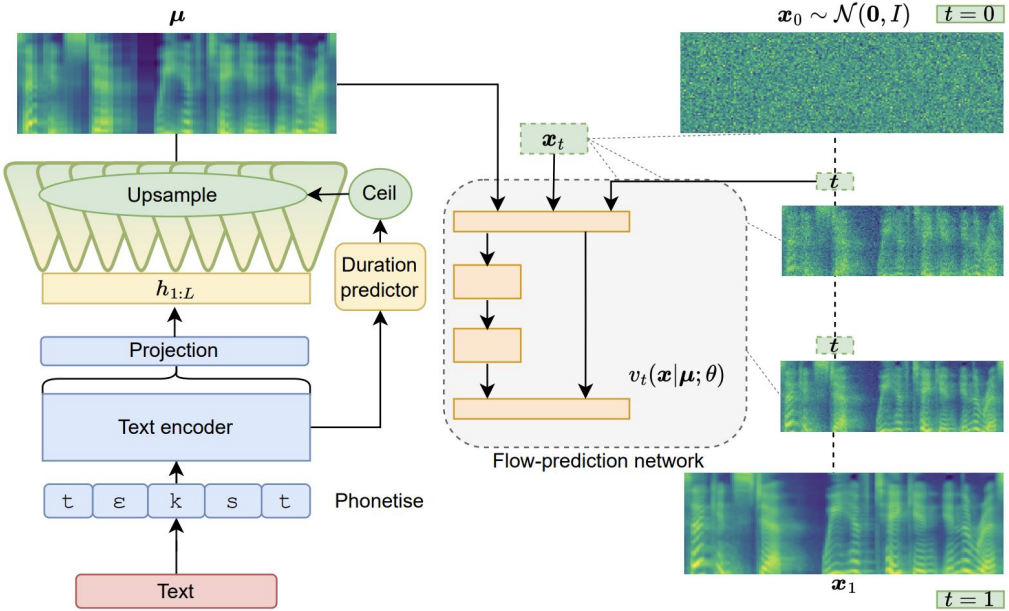
1. Sample $t \sim U(0, 1)$
2. Sample $(x_0, x_1) \sim \pi^*$
3. Compute $x_t = (1 - t)x_0 + tx_1$
4. Compute the target vector $v = x_1 - x_0$
5. Update θ to minimize $\|u_\theta(x_t, t) - v\|^2$

Sampling:

1. Sample $x_0 \sim p_0$
2. Solve the ODE: $\frac{d}{dt}x_t = u_\theta(x_t, t)$, with $x(0) = x_0$, from $t = 0$ to $t = 1$.

Matcha-TTS [2023]

- Model structure similar to GradTTS
- OT-CFM objective for mel-spec U-Net decoder
- U-Net decoder consists of Diffusion Transformer blocks
- Transformers with RoPE in the encoder, reducing RAM use and generalising to longer sequences
- Trained on LJspeech dataset
- 10 steps on inference



Condition	Params.	RAM	RTF ($\mu\pm\sigma$)	WER	MOS
VOC	13.9M	-	0.001±0.001	1.97	4.13±0.09
FS2	41.2M	6.0	**0.010**±0.004	4.18	3.29±0.09
VITS	36.3M	12.4	0.074±0.083	2.52	3.71±0.08
GRAD-10	**14.8M**	7.8	0.049±0.013	3.44	3.49±0.08
GRAD-4	"	"	0.019±0.006	3.69	3.20±0.09
GCFM-4	"	"	0.019±0.004	2.70	3.57±0.08
MAT-10	18.2M	**4.8**	0.038±0.019	**2.09**	**3.84**±0.08
MAT-4	"	"	0.019±0.008	2.15	3.77±0.07
MAT-2	"	"	0.015±0.006	2.34	3.65±0.08

Results on LJSpeech data

The Era of In-Context Learning

1. Key Enablers
    - Massive-Scale Data
    - Neural Audio Codecs
    - Architecture Breakthroughs:
        - Autoregressive Transformers: Natural sequence modeling for audio token prediction
        - Masked Reconstruction (BERT for text, MaskGIT for images)
        - Diffusion Models
2. Revolutionary Capabilities:
    - Multiple tasks with one architecture
    - Zero-Shot Adaptation: No fine-tuning needed for new voices/styles
    - Prompt-Based Control
3. Paradigm Shift: From specialized single-task models to general-purpose models that learns from context

Modern Paradigms of TTS models: Discrete approach

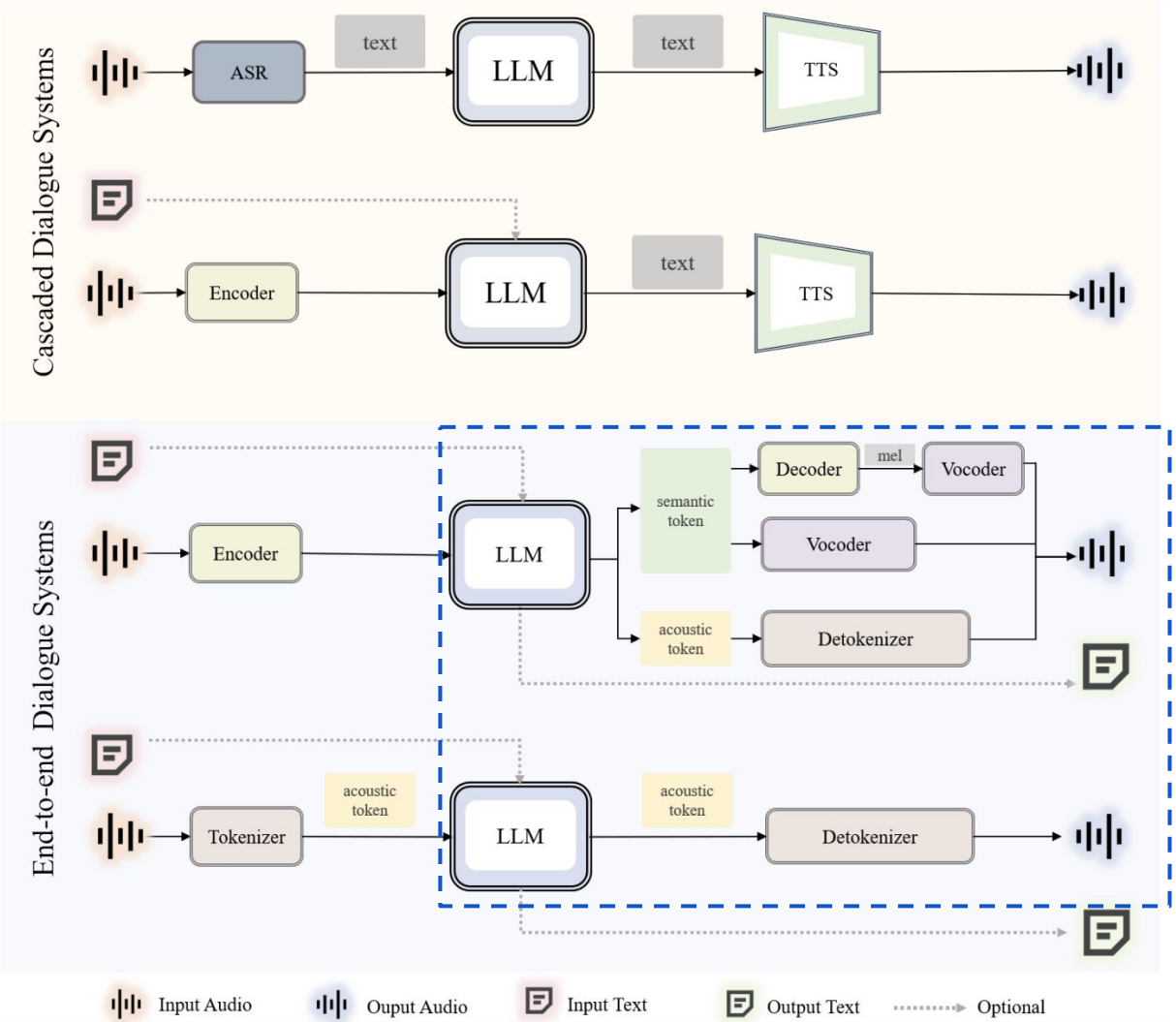
Core Idea: Treat speech synthesis as discrete sequence modeling

Architectures:

- AR: Pure Autoregressive (aka next-token-prediction)
- NAR: Non-autoregressive Masked Modeling (BERT for text, MaskGIT for images)
- Hybrid AR+NAR approaches

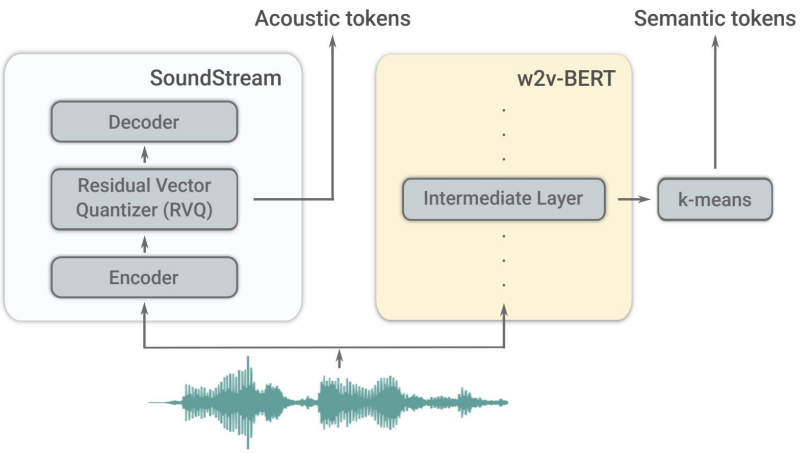
Advantages:

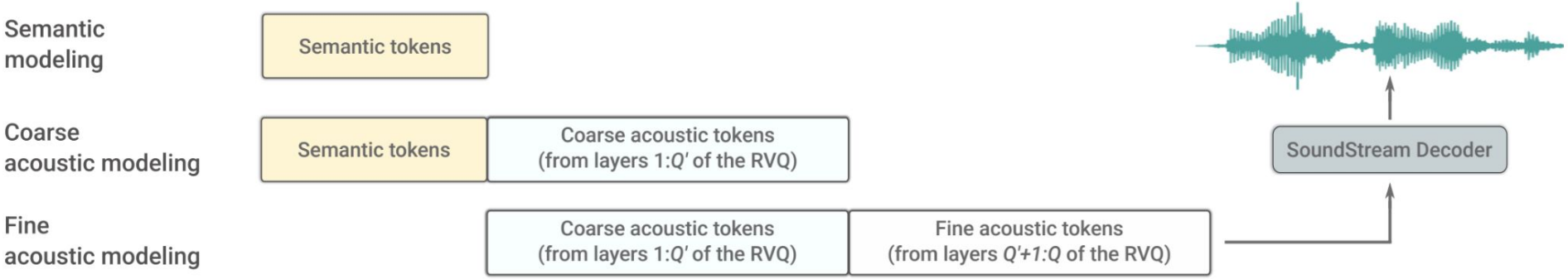
- Leverages LLM capabilities
- Strong in-context learning
- Efficient sequence generation


Cascaded Dialogue Systems
ASR
text
LLM
text
TTS
Encoder
LLM
text
TTS
End-to-end Dialogue Systems
Encoder
LLM
semantic token
Decoder
mel
Vocoder
Vocoder
acoustic token
Detokenizer
Tokenizer
acoustic token
LLM
acoustic token
Detokenizer
Input Audio
Ouput Audio
Input Text
Output Text
Optional

AudioLM [2022]

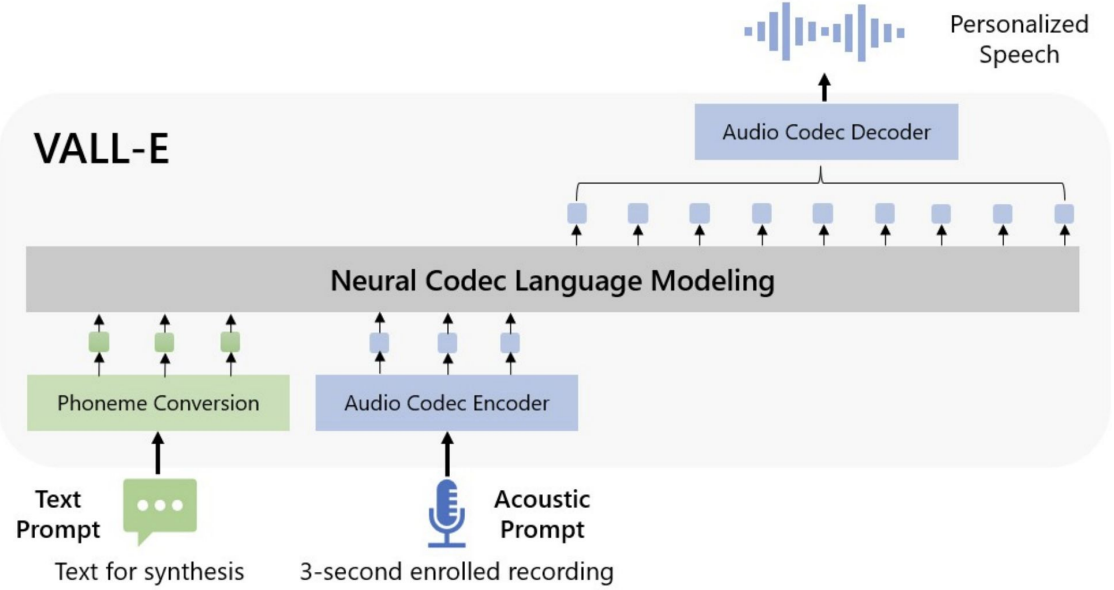
- 3-stage model to predict semantic and acoustic (coarse and fine) features:
  - 1st stage: predict semantic tokens obtained with k-means from SSL model
  - 2nd stage: predict coarse acoustic tokens from SoundStream codec
  - 3d stage: predict fine acoustic tokens from SoundStream codec
- On 2nd and 3d stage acoustic tokens flattened into single sequences
- Trained on 60k hours LibriLight dataset





VALL-E [Microsoft, 2023]

- Learns to generate EnCodec tokens with AR + NAR model
- AR predicts tokens from 1st codebook
- NAR learns to model the j-th codebook tokens given the embeddings up to j-1 level
- Trained on 60k hours LibriLight dataset

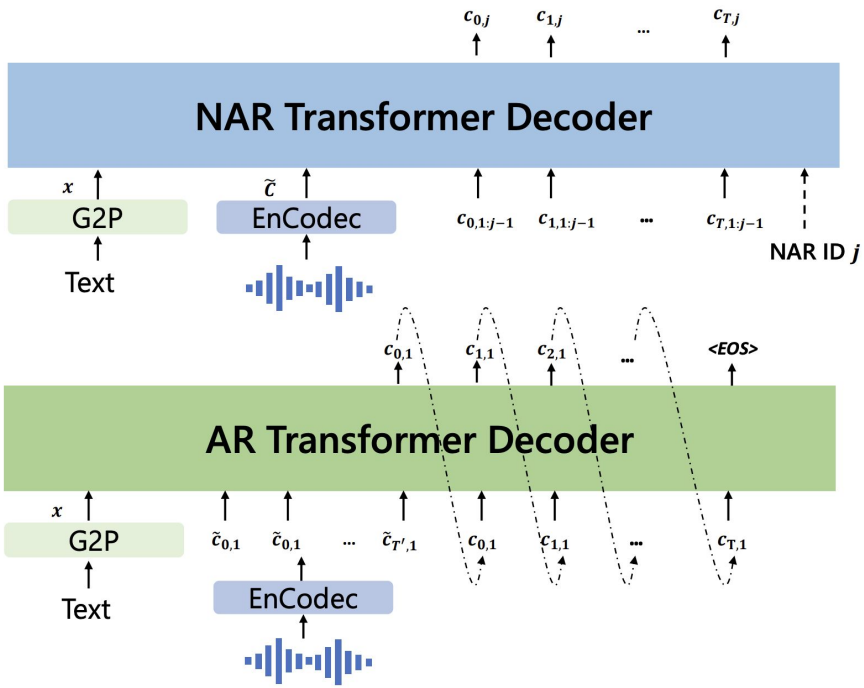


VALL-E [Microsoft, 2023]

- Learns to generate EnCodec tokens with AR + NAR model
- AR predicts tokens from 1st codebook
- NAR learns to model the j-th codebook tokens given the embeddings up to j-1 level
- Trained on 60k hours LibriLight dataset

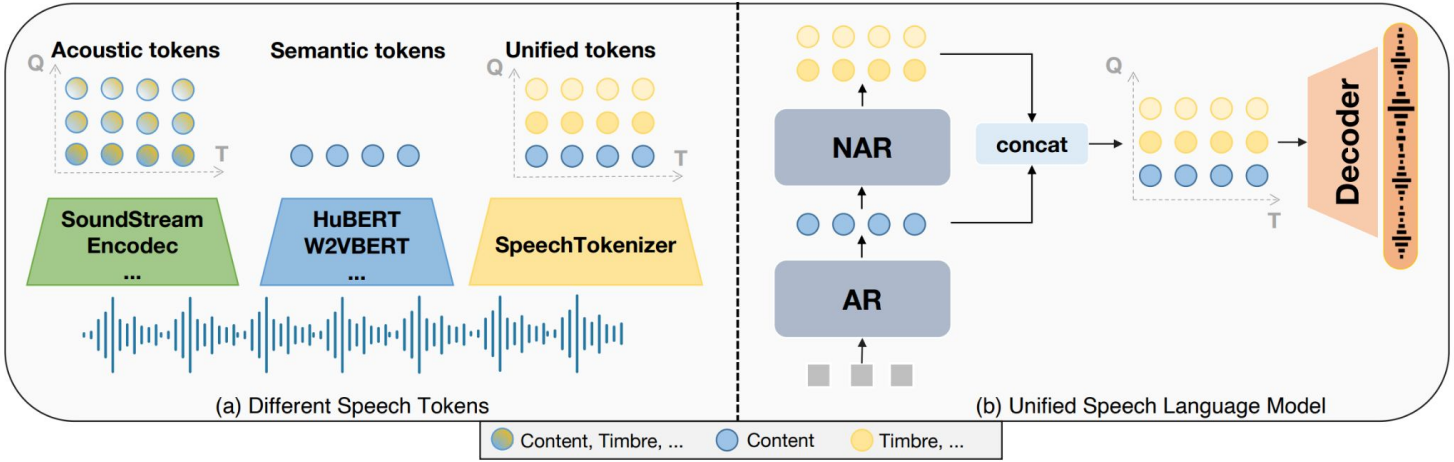
$$\mathcal{L}_{AR} = -\log \prod_{t=0}^{T} p(\mathbf{c}_1^t | \mathbf{c}_1^{<t}, \mathbf{u}; \theta_{AR})$$

$$\mathcal{L}_{NAR} = -\log \prod_{i=2}^{8} p(\mathbf{c}_i | \mathbf{c}_{<i}, \hat{\mathbf{C}}, \mathbf{u}; \theta_{NAR})$$



USLM [2023]

- Same idea as in VALL-E but with SpeechTokenizer codec
- In SpeechTokenizer semantic information distilled from SSL model (HuBERT) into the first RVQ codebook



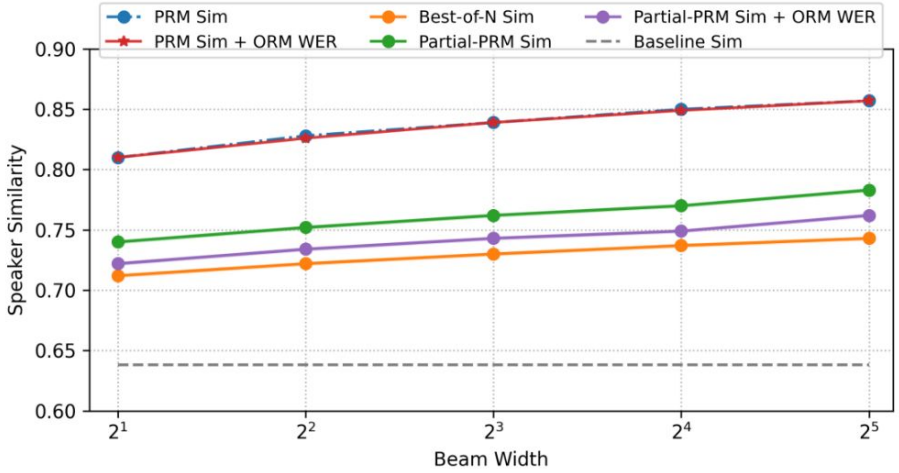
	Accurate Content	High-quality Speech	Single Tokenzier
Semantic LM	√	×	√
Acoustic LM	×	√	√
Hierarchical LM	√	√	×
USLM (ours)	√	√	√

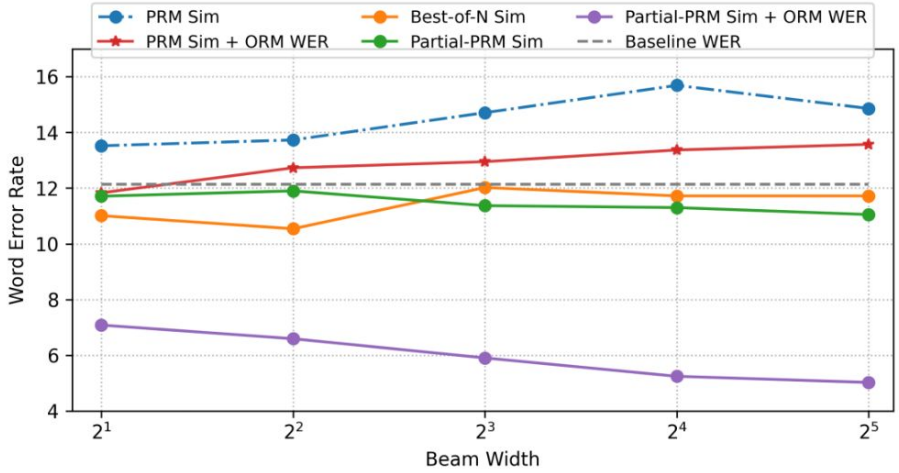
Llasa [2025]

- X-Codec2 single codebook speech tokenizer
- Utilize Output Reward Models and Process Reward Models to improve speaker similarity and WER
- Similarity verifier is a WavLM-finetuned speaker verification model
- Partial PRM strategy: PRM only during the first n seconds, then switching to ORM
- Substituting the later ORM step with a WER-based verifier simultaneously boosts both SIM and WER

Llasa [2025]

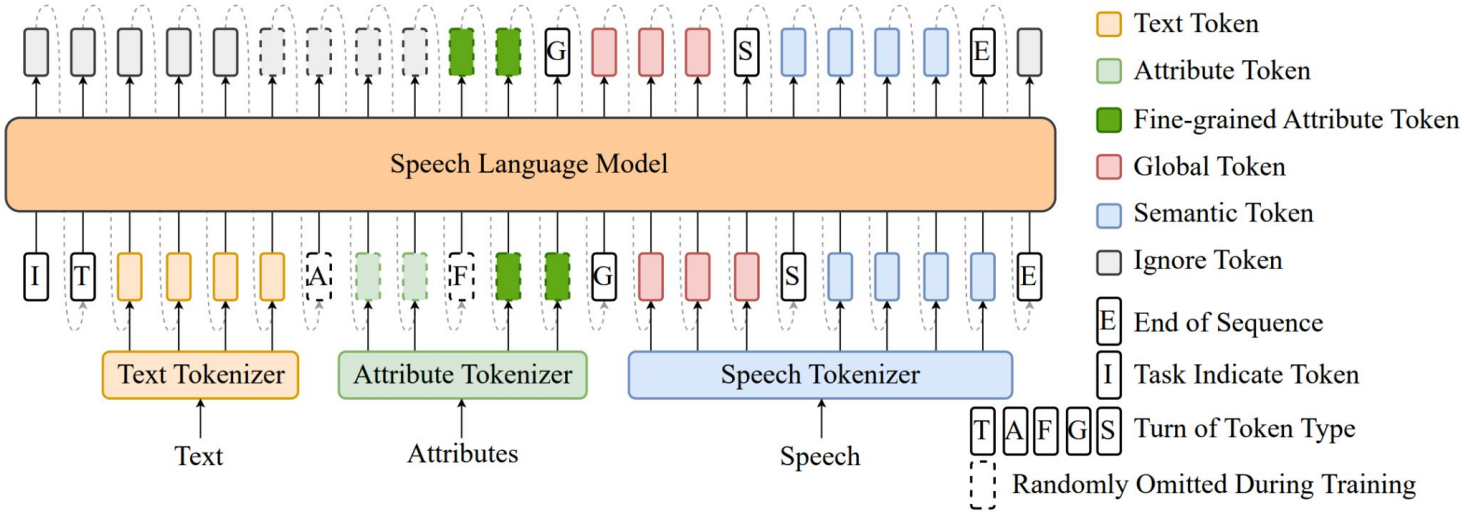
- X-Codec2 single codebook speech tokenizer
- Utilize Output Reward Models and Process Reward Models to improve speaker similarity and WER
- Similarity verifier is a WavLM-finetuned speaker verification model
- Partial PRM strategy: PRM only during the first n seconds, then switching to ORM
- Substituting the later ORM step with a WER-based verifier simultaneously boosts both SIM and WER





Spark-TTS [2025]

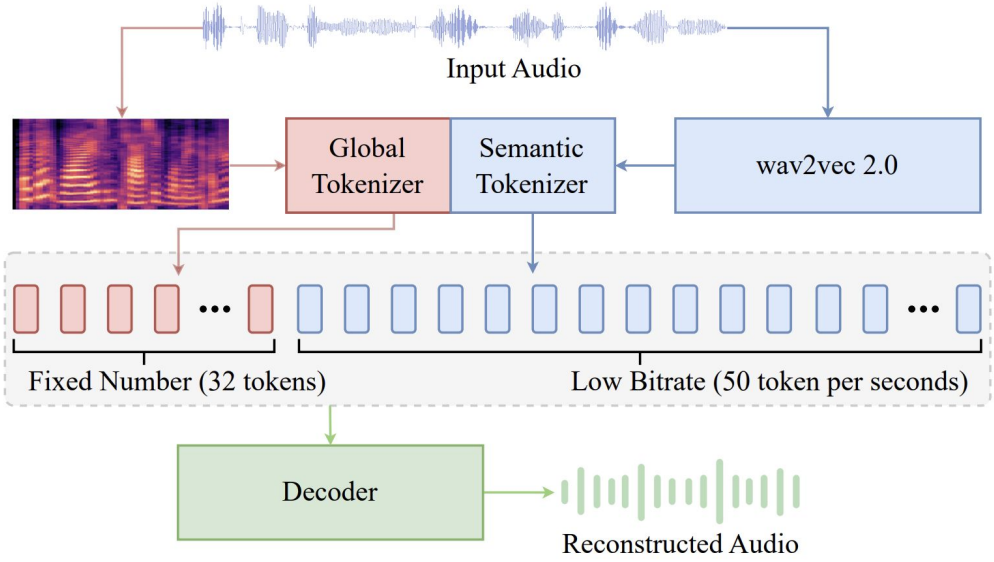
- Custom BiCodec: decoupled semantic and global tokens
- Single-stream sequence modeling
- Employ the pre-trained textual LLM Qwen2.5-0.5B as the backbone of the speech language model
- VoxBox training set



Spark-TTS [2025]

- Custom BiCodec: decoupled semantic and global tokens
- Single-stream sequence modeling
- Employ the pre-trained textual LLM Qwen2.5-0.5B as the backbone of the speech language model
- VoxBox training set

$$
\begin{aligned}
&\boldsymbol{z} = E_s(F(\boldsymbol{x})), \boldsymbol{g} = E_g(\mathrm{Mel}(\boldsymbol{x})), \\
&\boldsymbol{g}_f = \mathrm{CrossAttention}(\boldsymbol{g}, \boldsymbol{h}), \\
&\boldsymbol{z}_q = Q_s(\boldsymbol{z}), \boldsymbol{g}_q = Q_g(\boldsymbol{g}_f), \\
&\hat{\boldsymbol{x}} = G(\boldsymbol{z}_q, A_g(\boldsymbol{g}_q)),
\end{aligned}
$$

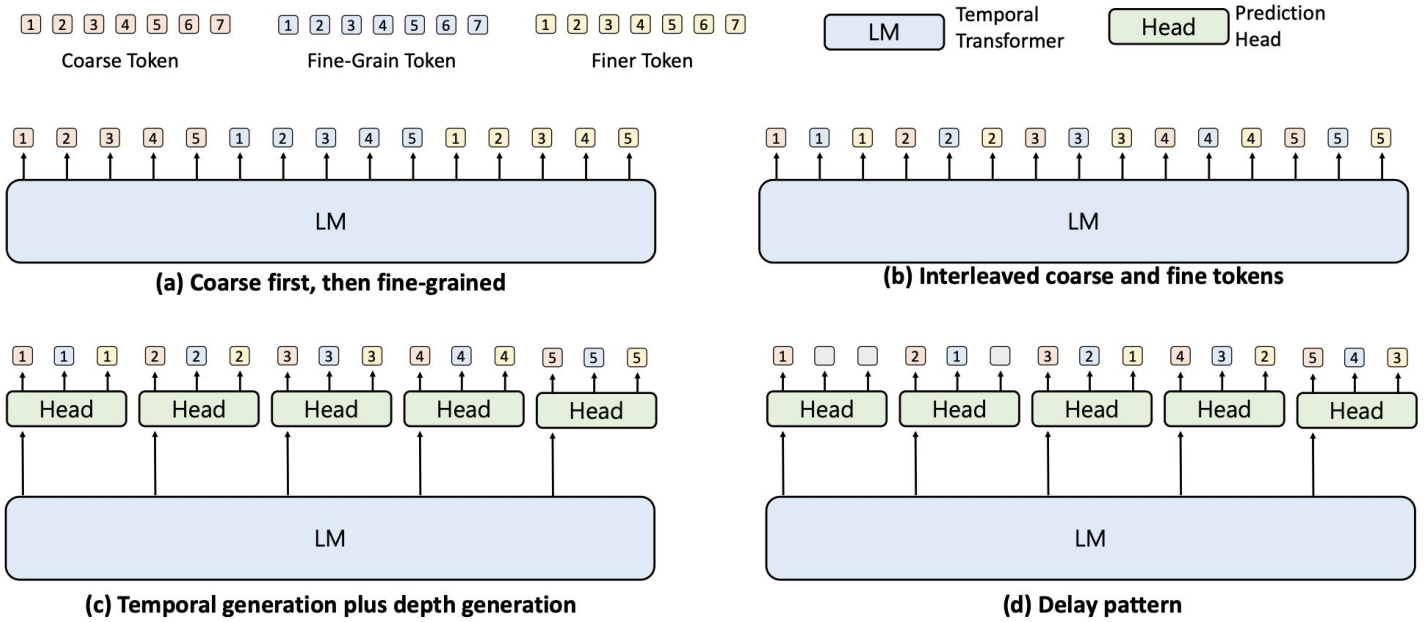


RVQ tokens decoding strategies

1. Generate one by one in a linear fashion (very long, too many steps, especially if there are multiple layers)

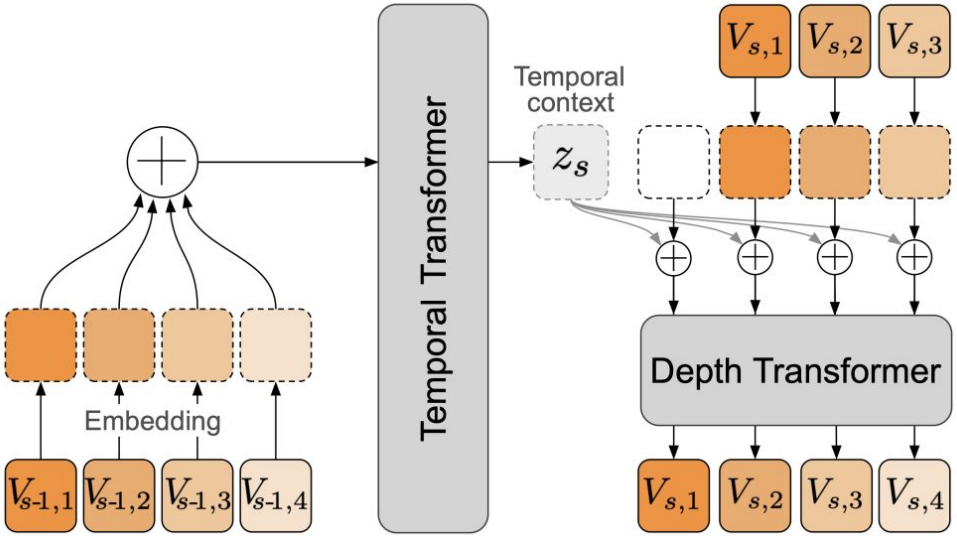
2. Generate in parallel (lower quality, the tokens are dependent on each other)

3. Stacked delay (a compromise between sequential and parallel generation)



RQ-Transformer and Depth-Transformer

- Idea from paper “Autoregressive Image Generation using Residual Quantization”
- RQ-Transformer breaks down a flattened sequence of length K · S into S timesteps for a large Temporal Transformer which produces a context embedding used to condition a smaller Depth Transformer over K steps



Acoustic Delay	Semantic Token Weight	Depthwise Parametrization	Inner Monologue	Transcript NLL (↓)	Transcript Length (↑)
[0, 0, 0, 0, 0, 0, 0, 0]	1.0	✓		4.36	486
[0, 1, 1, 1, 1, 1, 1, 1]	1.0	✓		4.12	529
[0, 2, 2, 2, 2, 2, 2, 2]	1.0	✓		4.09	519
[0, 2, 2, 2, 2, 2, 2, 2]	100.0			3.75	538
[0, 2, 2, 2, 2, 2, 2, 2]	100.0	✓		3.65	602
[0, 2, 2, 2, 2, 2, 2, 2]	100.0	✓	✓	**2.77**	**1920**

FishSpeech [FishAudio, 2024]

- Custom acoustic speech tokenizer with Grouped FSQ quantizer (RVQ analogue)
- Depth Transformer for speech tokens decoding from last hidden state of LLM
- 3 stage of training: pretraining, SFT, DPO
- 4B Llama-style backbone
- Dataset contains 720k hours of speech across different languages:
  - 300k hours each of English and Mandarin Chinese
  - 20,000 hours each of other language families: Germanic (German), Romance (French, Italian), East Asian (Japanese, Korean), and Semitic (Arabic)

Speech Generation as Masked Reconstruction task

Core paradigm:

- Inspired by BERT: Extends masked language modeling to speech domain
- Training Objective: Reconstruct masked portions of audio from context

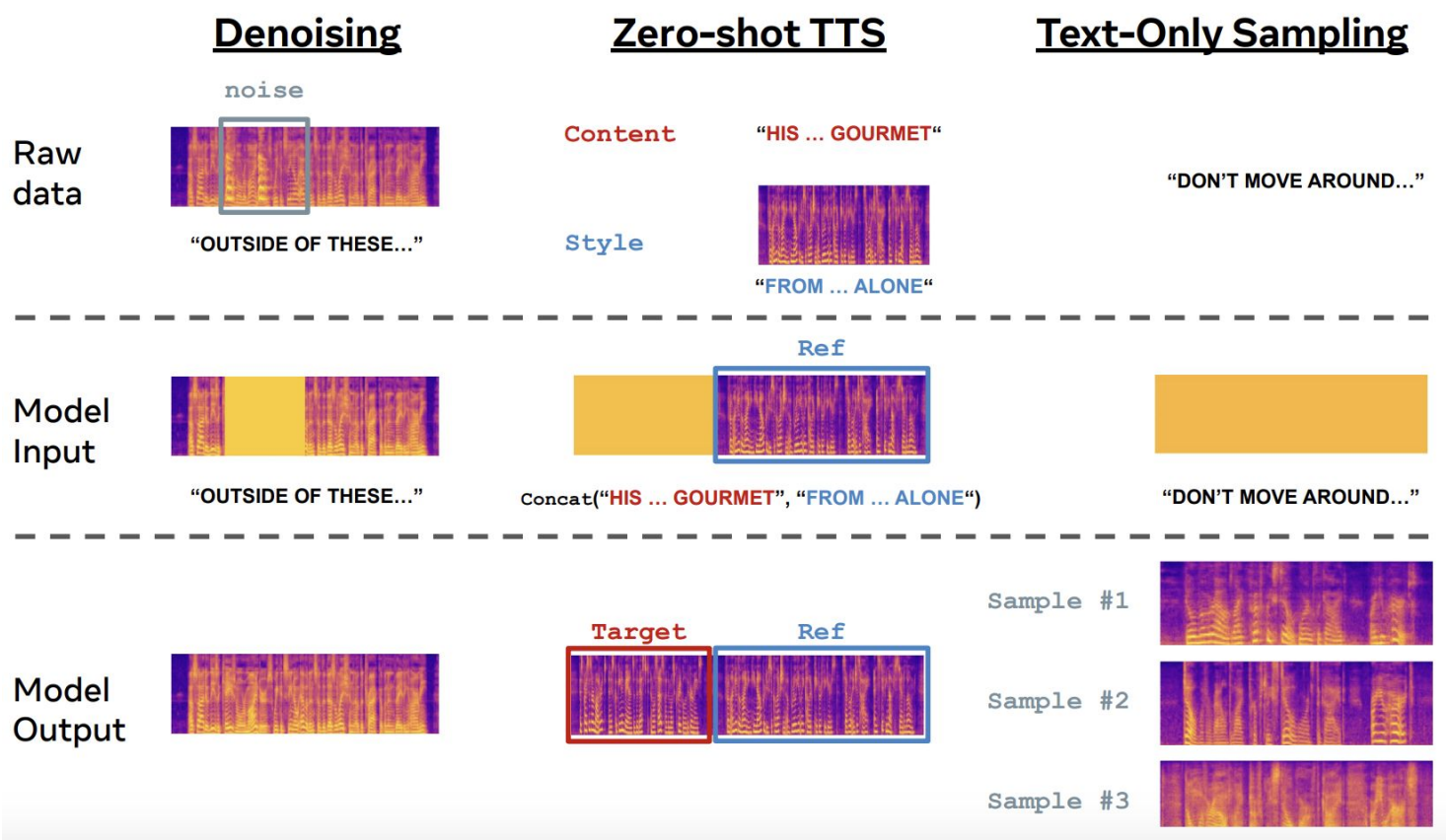
Advantages:

- Bidirectional Context: Leverages past and future information
- Efficient Parallel Generation: Non-autoregressive inference
- Strong Zero-Shot Capabilities: Natural in-context learning
- Unified Framework: Handles TTS, editing, and continuation tasks

Drawbacks:

- Bidirectional inference is inapplicable for long sequences and high-load systems

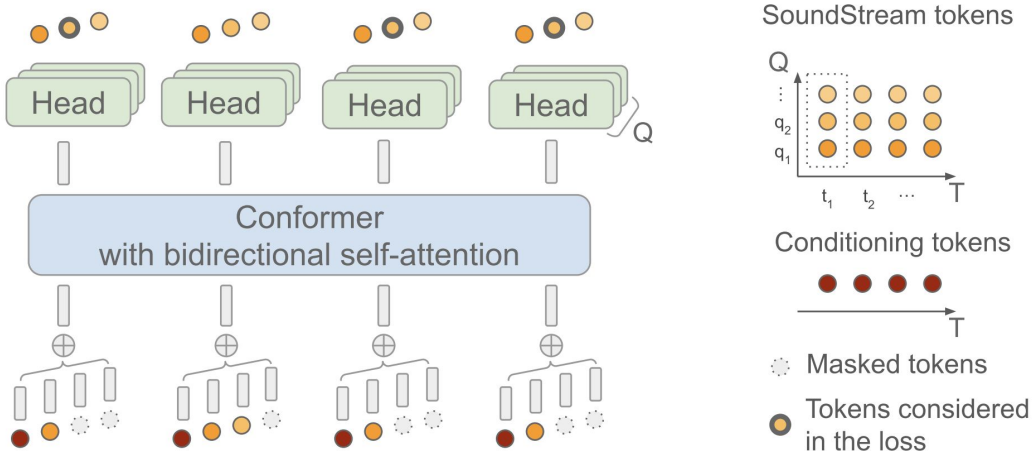
Speech Generation as Masked Reconstruction task



SoundStorm [Google Research, 2023]

- Learns to generate SoundStream tokens with AR + NAR model
- Idea from MaskGIT model for image generation
- Trained on 60k hours LibriLight dataset
- Masking algorithm on training stage:
  - Prompt delimiter timestep t ~ U{0, T −1}
  - Current RVQ level q ~ U{1, Q}
  - Sample the mask M for level q
  - Mask the selected non-prompt tokens at level q and all non-prompt tokens at finer RVQ levels

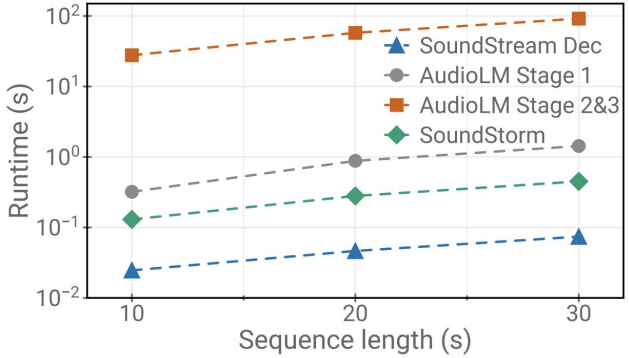
- Iterative Parallel Decoding on inference stage

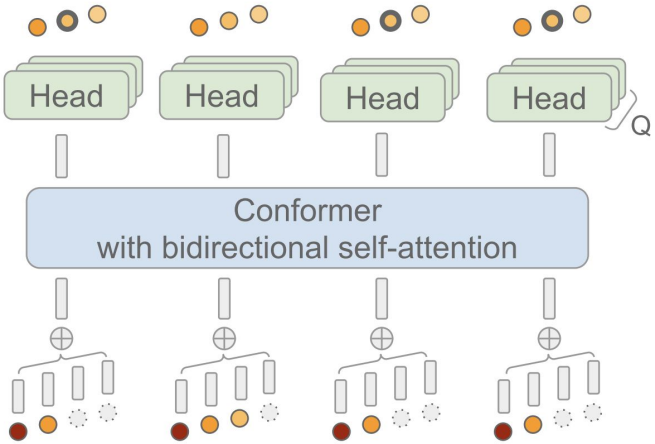


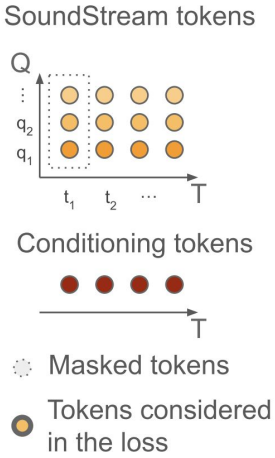
SoundStorm [Google Research, 2023]

- Learns to generate SoundStream tokens with AR + NAR model
- Idea from MaskGIT model for image generation
- Trained on 60k hours LibriLight dataset
- Masking algorithm on training stage:
  - Prompt delimiter timestep t ~ U{0, T −1}
  - Current RVQ level q ~ U{1, Q}
  - Sample the mask M for level q
  - Mask the selected non-prompt tokens at level q and all non-prompt tokens at finer RVQ levels

- Iterative Parallel Decoding on inference stage

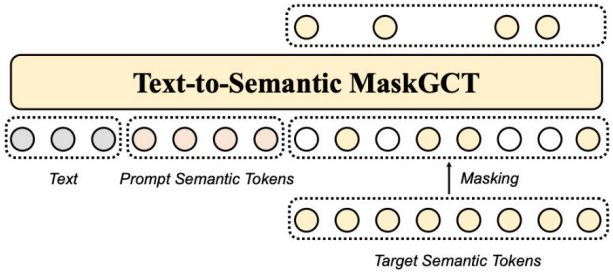


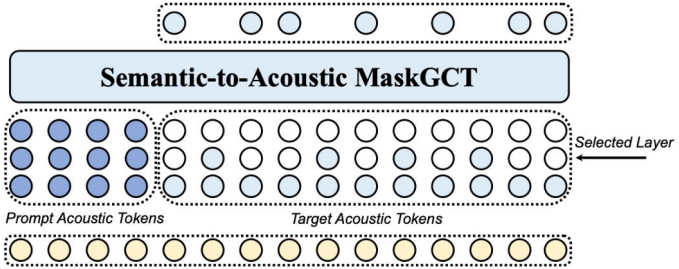


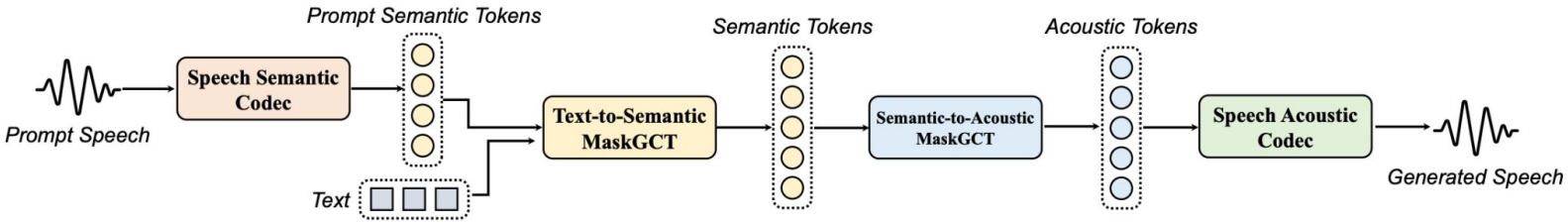


Mask-GCT [2024]

- Separate semantic and acoustic tokenizers
- Improves SoundStorm by introducing text-to-semantic stage
- Trained on 100K hours Emilia dataset (English+Chinese)
- 50 steps as the default total inference steps

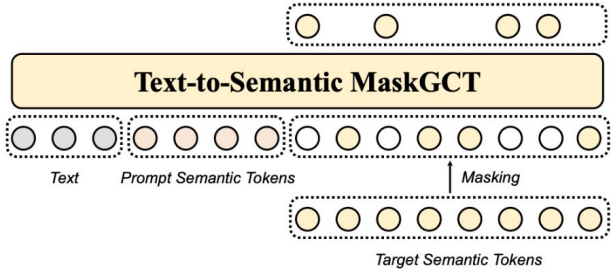


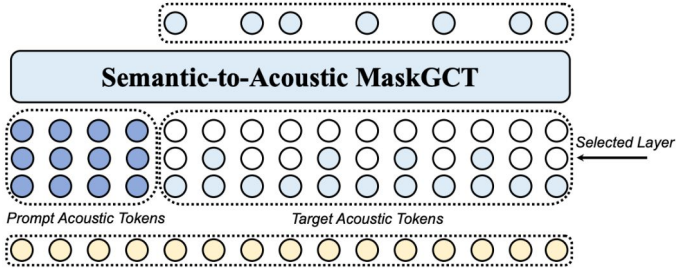




Mask-GCT [2024]

- Separate semantic and acoustic tokenizers
- Improves SoundStorm by introducing text-to-semantic stage
- Trained on 100K hours Emilia dataset (English+Chinese)
- 50 steps as the default total inference steps





System	Model	Rep.	Imp. Dur.	Len. Ctrl.	ZS TTS	CL TTS	Dubbing	Edit
VALL-E	Autoregressive	A. Tokens	✓	✗	✓	✗	✗	✗
NaturalSpeech 2	Diffusion	A. Features	✗	✗	✓	✗	✗	✗
VoiceBox	Diffusion	A. Features	✓	✗	✓	✓	✗	✓
VoiceCraft	Autoregressive	A. Tokens	✓	✗	✓	✗	✗	✓
NaturalSpeech 3	Masked Generative	F. Tokens	✗	✗	✓	✗	✗	✓
MaskGCT	Masked Generative	S.&A. Tokens	✓	✓	✓	✓	✓	✓

Modern Paradigms of TTS models: Continuous approach

Core idea: Direct high-fidelity audio generation

Method: Model mel-spectrograms using diffusion/flow matching

Architectures:

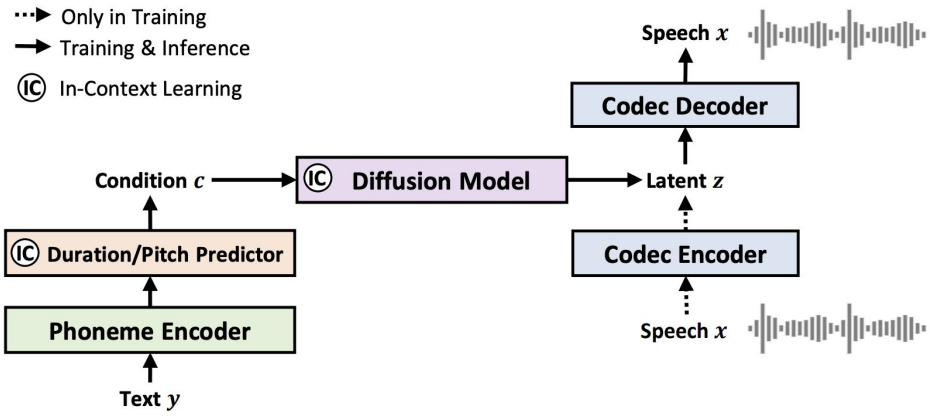
- Diffusion Models and Flow Matching (Conv U-Net, DiT)

Advantages:

- Superior audio quality
- Absent of lossy quantization
- Native editing
- Natural in-context learning

Natural Speech 2 [Microsoft Research, 2023]

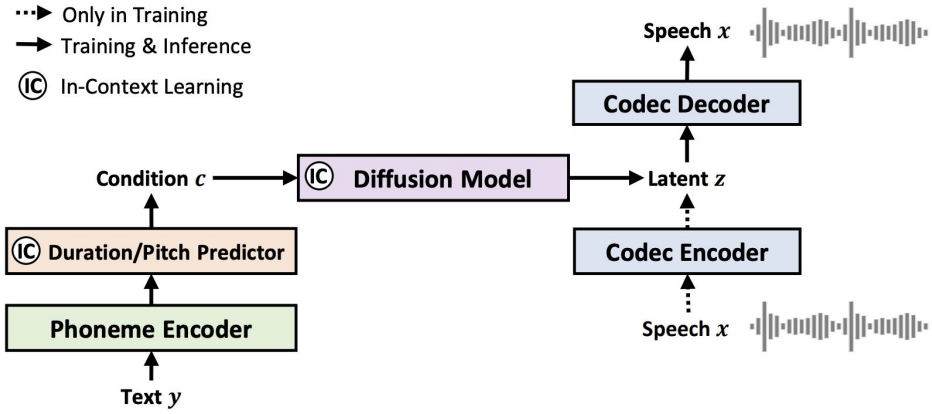
- Learns to generate EnCodec features
- Short speech prompt is used for conditioning timbre and prosody
- Trained on large 44k hours English subset of Multilingual LibriSpeech
- 150 diffusion steps on inference
- Can extend the styles beyond speech (e.g., singing voice)
- Can be extended to Voice Conversion and Speech Enhancement under the source-aware and target-aware diffusion processes

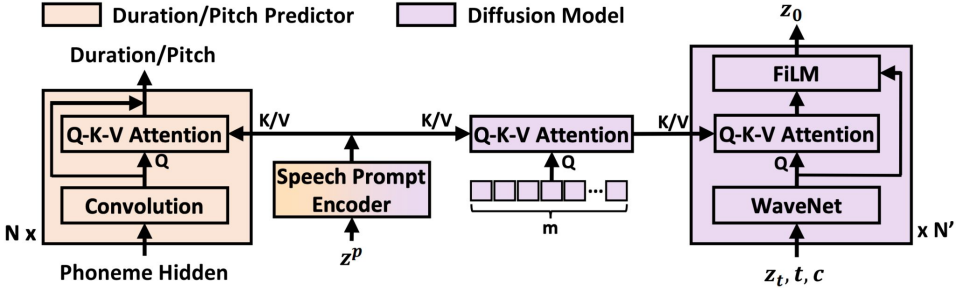


The Dilemma in Previous Systems	Single Token (VQ)	Multiple Tokens (RVQ)
Waveform Reconstruction (Discrete Audio Codec)	Hard	Easy
Token Generation (Autoregressive Language Model)	Easy	Hard

Natural Speech 2 [Microsoft Research, 2023]

- Learns to generate EnCodec features
- Short speech prompt is used for conditioning timbre and prosody
- Trained on large 44k hours English subset of Multilingual LibriSpeech
- 150 diffusion steps on inference
- Can extend the styles beyond speech (e.g., singing voice)
- Can be extended to Voice Conversion and Speech Enhancement under the source-aware and target-aware diffusion processes

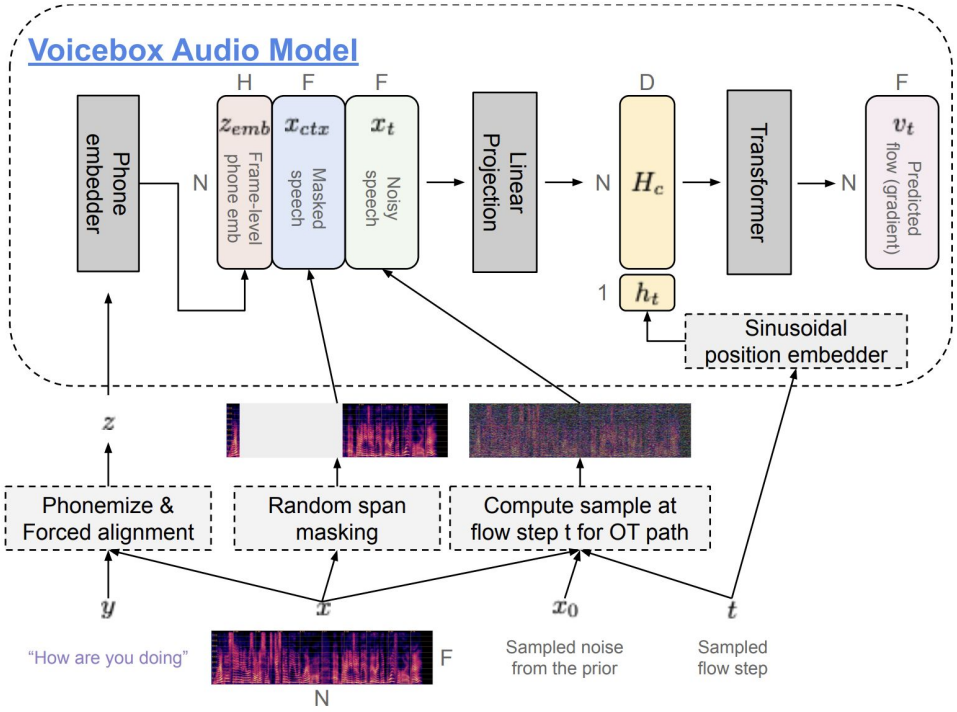




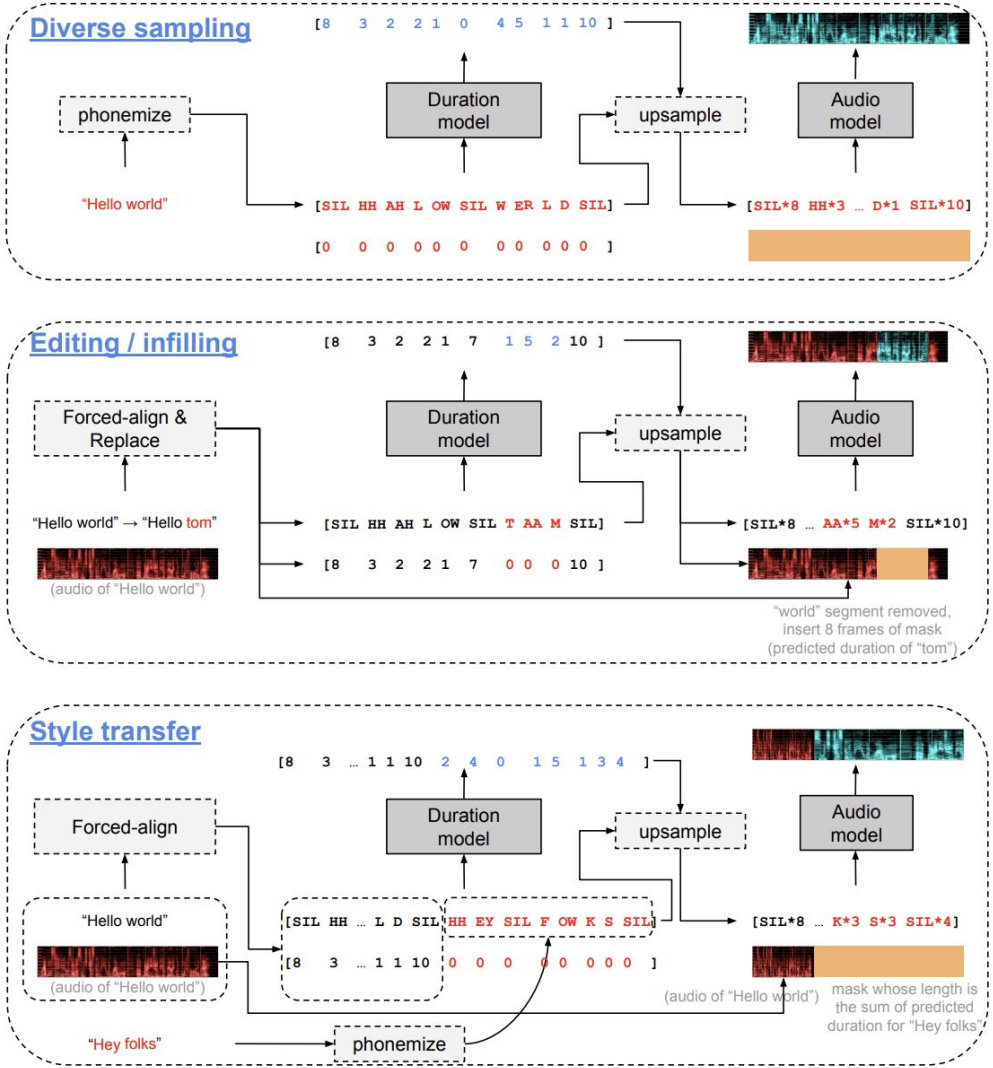
$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,t}[||\hat{z}_0 - z_0||_2^2 + ||\Sigma_t^{-1}(\rho(\hat{z}_0, t) - z_t) - \nabla \log p_t(z_t)||_2^2 + \lambda_{ce-rvq}\mathcal{L}_{\text{ce-rvq}}]$$

VoiceBox [FAIR, 2023]

- One of the first tts model solving text-guided speech infilling task
- First model that can perform high-quality cross-lingual zero-shot TTS across six languages
- Capable to solve tasks in zero-shot manner due to in-context learning
- English-only model trained on 60K hours audiobooks and the multilingual model on 50K hours
- 32 ODE solver steps on inference

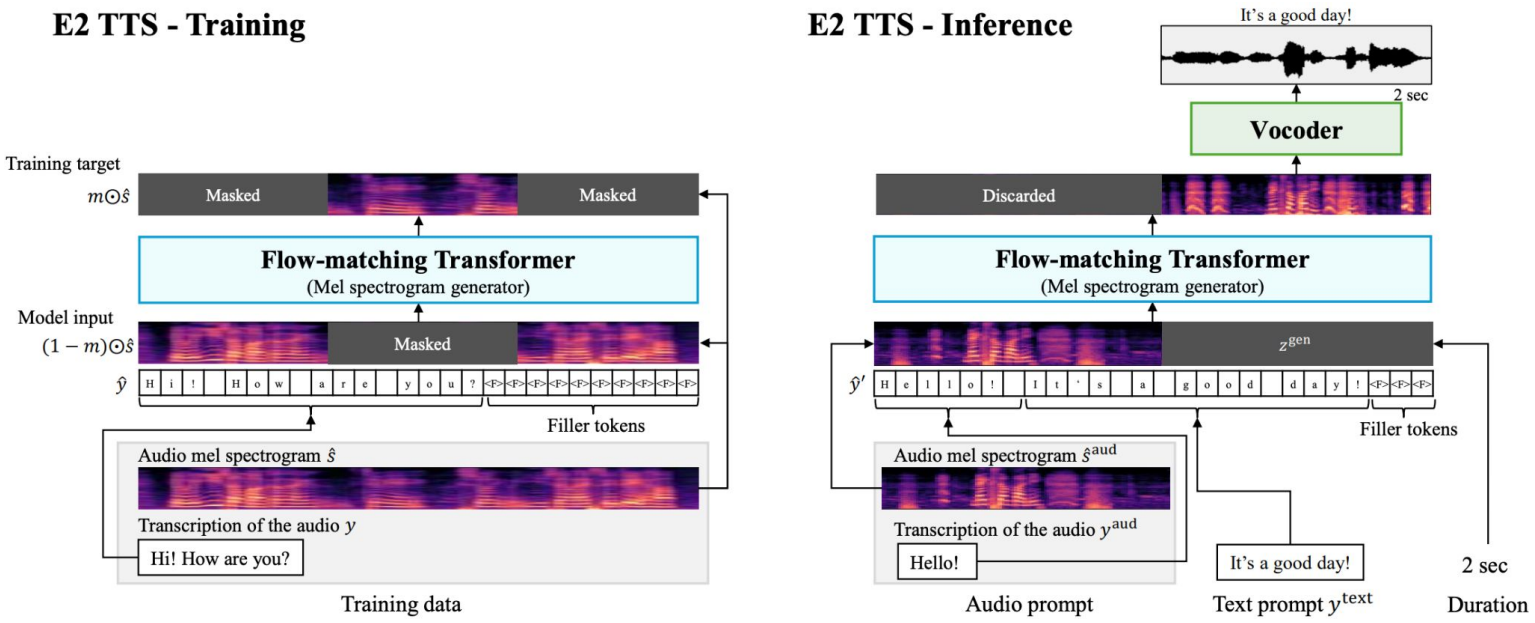


VoiceBox [FAIR, 2023]



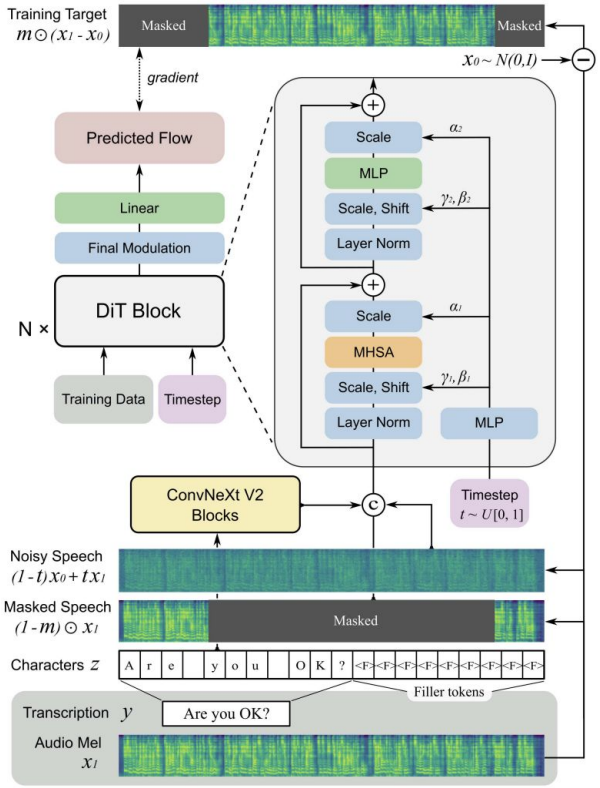
E2-TTS [Microsoft, 2024]

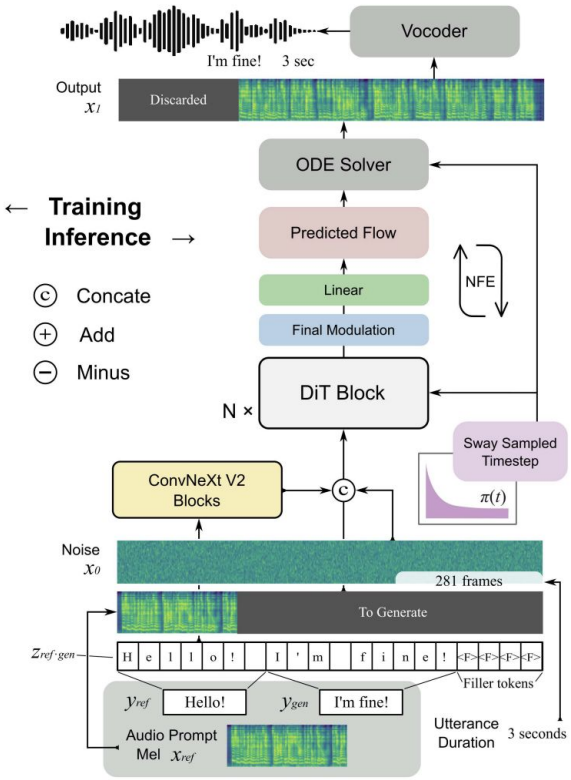
- First NAR approach without explicit duration modeling
- No grapheme-to-phoneme modeling
- Use 50k hours from LibriLight and 200k hours of in-house data for ablation
- 32 ODE solver steps on inference
- Trained on 100K hours Emilia dataset (English+Chinese)



F5-TTS [2024]

- E2 TTS converges slowly and isn't robust
- F5 proposes an inference-time Sway Sampling strategy, which significantly improves model's performance and efficiency
- CFM focus more on sketching the contours of speech in the early stage from pure noise and later focus more on the embellishment of fine-grained details
- Inference RTF of 0.15

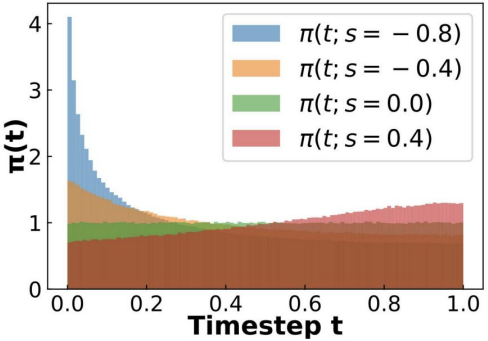


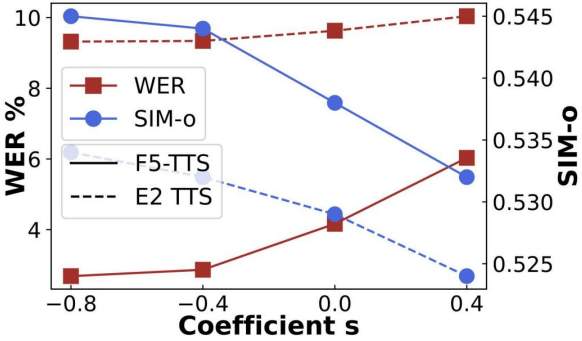


F5-TTS [2024]

- E2 TTS converges slowly and isn't robust
- F5 proposes an inference-time Sway Sampling strategy, which significantly improves model's performance and efficiency
- CFM focus more on sketching the contours of speech in the early stage from pure noise and later focus more on the embellishment of fine-grained details
- Inference RTF of 0.15

$$f_{sway}(u; s) = u + s \cdot (\cos(\frac{\pi}{2}u) - 1 + u)$$





Modern Paradigms of TTS models: Hybrid Approach

Core Architecture:

- Semantic planning: LLM generates high-level semantic tokens from text
- Acoustic refinement: Diffusion/Flow models produce fine-grained acoustic features

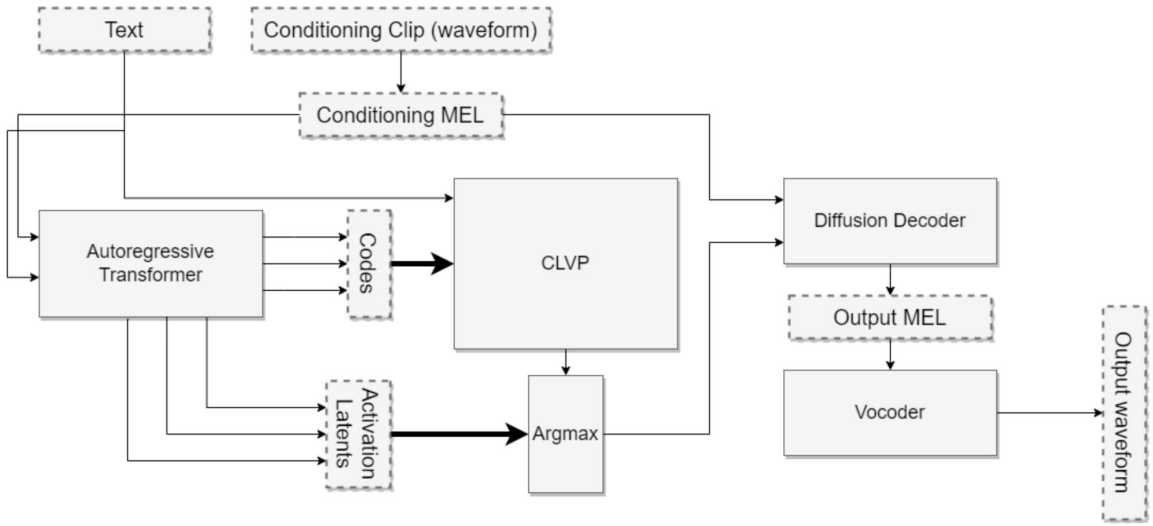
Advantages:

- LLM context understanding + diffusion audio quality
- Optimized quality-speed tradeoff, supports streaming
- Fine-grained adjustment of speaking style and emotion
- Efficient Scaling

Emerging Trend: Hybrid architectures are becoming the dominant paradigm

TorToise TTS [James Betker, 2023]

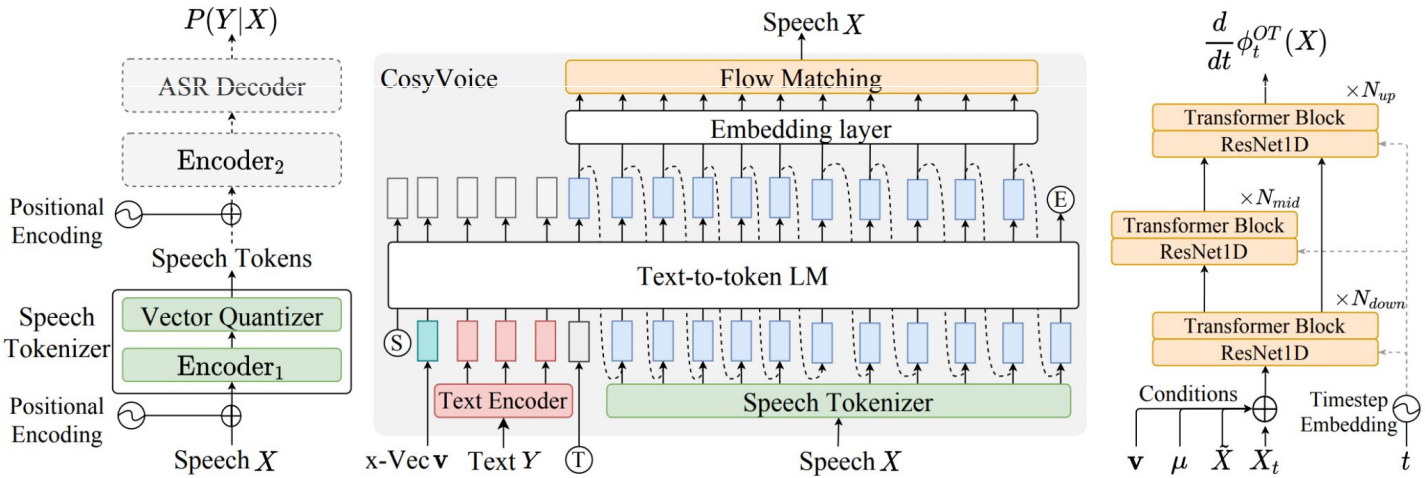
- AR decoder predicts a probability distribution for speech tokens
- Denoising Diffusion can convert speech tokens back into speech spectrograms
- **Key insight**: “For the majority of the training, Diffusion converts discrete speech codes into mel-spectrograms. [...], I **fine-tune the DDPM on the latent space** which is pulled from the AR model outputs [...]. By fine-tuning on this latent space, we improve the efficiency of the downstream diffusion model.”
- This **fine-tuning** is one of the **greatest contributors to model output quality**.
- The least important part: contrastive model similar to CLIP is used to rank outputs of the AR decoder



CosyVoice [Alibaba, 2024]

- Adapts the idea from TorToiseTTS
- Vector-quantized hidden representations as semantic vocabulary
- Data subset labeled with instructions opens the door for instruction synthesis
- CFM conditions on speaker embedding

Type	Duration (hr)
Speaker Identity	101
Speaking Style	407
Fine-grained Paralinguistics	48



(a) Supervised speech tokenizer (b) An overview of the proposed CosyVoice LM (c) Conditional flow matching

Language	Duration (hr)
ZH	130,000
EN	30,000
Yue	5,000
JP	4,600
KO	2,200

CosyVoice [Alibaba, 2024]

Speaker Identity
1. Selene 'Moonshade', is a **mysterious**, **elegant dancer** with a connection to the night. Her movements are both **mesmerizing** and **deadly**.<endofprompt>Hope is a good thing.
2. Theo 'Crimson', is a **fiery**, **passionate** rebel leader. Fights with fervor for justice, but struggles with **impulsiveness**.<endofprompt>You don't know about real loss.

Speaking Style
1. A **happy girl** with **high tone** and **quick speech**.<endofprompt>The sun is shining brightly today.
2. A **sad woman** with **normal tone** and **slow speaking speed**.<endofprompt>I failed my important exam.

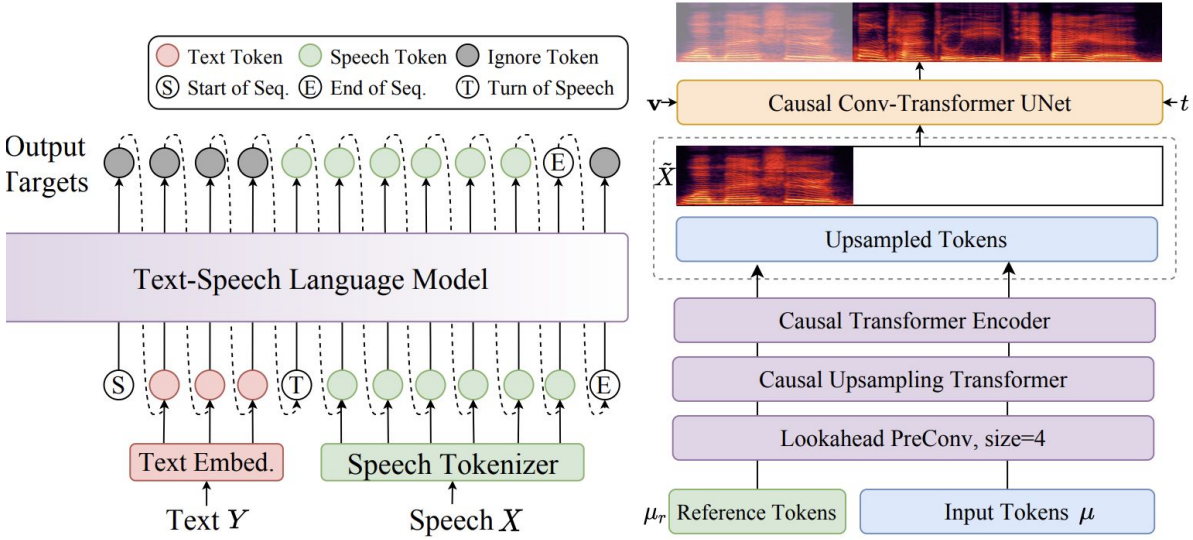
Fine-grained Paralinguistics
1. Well that's kind of scary **[laughter]**.
2. I don't think I over eat yeah **[breath]** and um I do exercise regularly.
3. Well that pretty much covers <**laughter**>**the subject**<**/laughter**> well thanks for calling me.
4. The team's <**strong**>**unity**<**/strong**> and <**strong**>**resilience**<**/strong**> helped them win the championship.

Model	Happy	Sad	Angry	Surprised	Fearful	Disgusted
CosyVoice-base	1.00±0.00	0.45±0.05	0.59±0.03	0.26±0.02	0.88±0.01	0.46±0.06
CosyVoice-instruct	1.00±0.00	0.98±0.02	0.83±0.04	0.64±0.03	0.87±0.03	0.93±0.02
w/o instruction	0.98±0.01	0.77±0.04	0.49±0.12	0.28±0.06	0.83±0.04	0.45±0.16

CosyVoice 2 [Alibaba, 2024]

- Initialize text-to-semantic model with pretrained LLM
- Chunk-aware Flow Matching technique for streaming inference
- WER and SS are used to distinguish preferred and rejected samples for DPO

Language	Duration (hours)
Chinese	130,000
English	30,000
Japanese	4,600
Korean	2,200



(b) Text-Speech Language Model (c) Chunk-aware Flow Matching

Why not use continuous latent representations?

Challenges with Discrete Token Modeling:

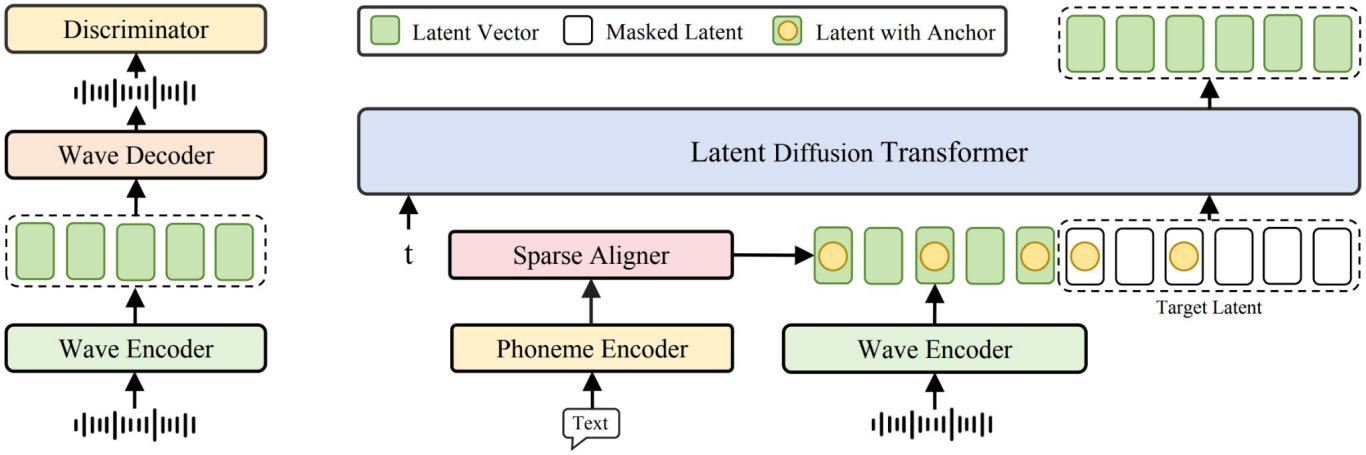
- Complex decoding strategies
- Quantization artifacts and information loss
- Separate modeling of semantic and acoustic tokens
- Discrete representations struggle with fine acoustic details

Advantages of Continuous Latent Spaces:

- Preserves subtle acoustic features and prosodic variations
- Enables interpolation and smooth transitions between speech characteristics
- Better capture of voice identity and emotional nuance
- Natural metric space

MegaTTS 3 [ByteDance, 2025]

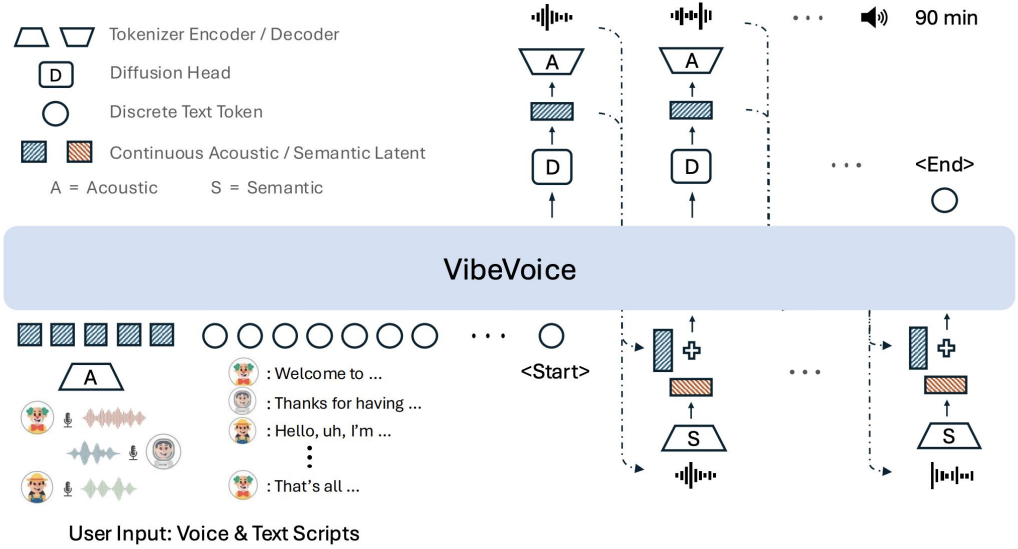
- Predict WavVAE latent representations instead of mel of audio tokens using Latent DiT
- Exploits pretrained WaveVAE
- Fully non-autoregressive
- Trained on 60k hours LibriLight dataset



(a) WaveVAE (b) Sparse-Aligned Diffusion Transformer

VibeVoice [Microsoft Research, 2025]

- VAE quantization-free acoustic tokenizer
- Semantic tokenizer tuned for ASR task
- Next-token-diffusion: Token-wise diffusion head predicts continuous latent representations from LLM latent
- The whole model is fully autoregressive
- LLM and diffusion head are trained in end-to-end manner with diffusion loss
- Extra-low frame rate (7.5Hz)
- Able to generate up to 90 minutes dialogue



Evaluations

VALL-E

	3s prompt	5s prompt	10s prompt
108 full speakers			
YourTTS*	0.357	0.377	0.394
VALL-E	0.382	0.423	**0.484**
GroundTruth	0.546	0.591	0.620
11 unseen speakers			
YourTTS	0.331	0.337	0.344
VALL-E	0.389	0.380	**0.414**
GroundTruth	0.528	0.556	0.586

Automatic evaluation of speaker similarity with 108 speakers on VCTK. *YourTTS has observed 97 speakers during training, while VALL-E observed none of them

	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

Human evaluation with 60 speakers on VCTK with 3-second enrolled recording for each

	SMOS	CMOS (v.s. VALL-E)
YourTTS	$3.45_{\pm 0.09}$	-0.12
VALL-E	$4.38_{\pm 0.10}$	0.00
GroundTruth	$4.5_{\pm 0.10}$	+0.17

Human evaluation with 40 speakers on LibriSpeech test-clean with 3-second enrolled recording for each

- speaker similarity measured as cosine distance between embeddings obtained with WavLM-TDNN-large finetuned for speaker verification if not stated opposite
- SMOS - similarity MOS
- CMOS - comparative MOS

USLM, Natural Speech 2, VoiceBox

Model	Tokenizer	Objective WER↓	Objective SIM↑	Subjective MOS↑	Subjective SMOS↑
Groundtruth		1.9	0.93	4.5	3.96
VALL-E	EnCodec	7.9	0.75	3.08	3.31
USLM	SpeechTokenizer	**6.5**	**0.84**	**3.63**	**3.45**

Results of zero-shot TTS on the VCTK dataset, which comprises 108 speakers with 3-second enrolled recording for each.

Model	**WER**	**SIM-o**	**SIM-r**	**QMOS**	**SMOS**
Ground truth	2.2	0.754	n/a	$3.98_{\pm 0.14}$	$4.01_{\pm 0.09}$
cross-sentence					
A3T	63.3	0.046	0.146	-	-
YourTTS	7.7	0.337	n/a	$3.27_{\pm 0.13}$	$3.19_{\pm 0.14}$
VALL-E	5.9	-	0.580	-	-
VB-En	1.9	0.662	0.681	$3.78_{\pm 0.10}$	$3.71_{\pm 0.11}$
continuation					
A3T	18.7	0.058	0.144	-	-
VALL-E	3.8	0.452*	0.508	-	-
VB-En ($\alpha = 0.7$)	2.0	0.593	0.616	-	-

English Zero-shot TTS results on filtered LS test-clean. SIM-o and SIM-r measure similarity to GT and reconstruction respectively

Setting	LibriSpeech	VCTK
Ground Truth	1.94	9.49
YourTTS	7.10	14.80
NaturalSpeech 2	**2.26**	**6.99**

WER results

Setting	LibriSpeech	VCTK
GroundTruth	3.33	3.86
YourTTS	2.03	2.43
NaturalSpeech 2	**3.28**	**3.20**

SMOS results

Setting	SMOS	CMOS
GroundTruth	4.09	-
VALL-E	3.53	−0.31
NaturalSpeech 2	**3.83**	**0.00**

Comparison on 8 samples from LibriSpeech and 8 samples from VCTK

E2-TTS, CosyVoice

ID	Model	Data (hours)	Init	WER↓	SIM-o↑
(GT)	Ground Truth	-	-	2.0	0.695
(B1)	VALL-E [8]	LL (60K)	Random	4.9	0.500
(B2)	NaturalSpeech 3 [15]	LL (60K)	Random	2.6	0.632
(B3)	Voicebox [18]†	LL (60K)	Random	2.1	0.658
(B4)	Voicebox [18]†	LH (50K)	Random	2.2	0.667
(B5)	Voicebox [18]†	LH (50K)	Pretrain [34]	2.2	0.695
(P1)	E2 TTS	LH (50K)	Random	2.0	0.675
(P2)	E2 TTS	LH (50K)	Pretrain [34]	**1.9**	**0.708**
(P3)	E2 TTS	PP (200K)	Random	**1.9**	0.707

Objective results for LibriSpeech-PC test-clean evaluation set

ID	Model	CMOS↑	SMOS↑
(GT)	Ground Truth	0.00	$3.91_{\pm 0.13}$
(B2)	NaturalSpeech 3 [15]	-0.98	$\mathbf{4.76}_{\pm 0.06}$
(B4)	Voicebox [18]†	-0.78	$4.73_{\pm 0.06}$
(P1)	E2 TTS	-0.14	$4.66_{\pm 0.07}$
(P2)	E2 TTS	**-0.05**	$4.65_{\pm 0.08}$
(P3)	E2 TTS	-0.18	$4.64_{\pm 0.08}$

Subjective results for LibriSpeech-PC test-clean evaluation set

Model	Text token	Speech Token	WER (%)	#INS+DEL	#SUB	SS
Original	-	-	3.01	66	200	69.67
VALL-E (Wang et al., 2023)	Phone	Encodec	18.70	342	1312	53.19
UniAudio (Yang et al., 2023)	Phone	Encodec	8.74	254	519	47.56
SpearTTS (Kharitonov et al., 2023)	Phone	Hubert	6.14	133	410	51.71
Exp-1-LibriTTS	Phone	Hubert	7.41	325	409	67.85
Exp-2-LibriTTS	Phone	S^3_{en}	5.05	122	325	67.85
Exp-3-LibriTTS	BPE_{en}	S^3_{en}	3.93	108	239	67.85
Exp-4-LibriTTS	BPE	S^3	4.76	134	287	65.94
Exp-4-Large-scale	BPE	S^3	**3.17**	**96**	**184**	**69.49**

Comparison of CosyVoice on the LibriTTS test-clean set

SoundStorm, Mask-GCT

	WER↓			CER↓			Audio quality↑			Voice preservation↑			Acoustic consistency↑		
	short	mid	long	short	mid	long	short	mid	long	short	mid	long	short	mid	long
Original SoundStream rec.	2.62	1.95	2.20	0.89	0.55	0.69	3.72	3.91	3.99	0.63	0.65	0.66	0.97	0.95	0.93
Without a speaker prompt															
AudioLM	4.65	3.59	4.79	2.15	1.57	2.30	3.93	4.04	4.08	–	–	–	–	–	–
SoundStorm	**3.48**	**2.55**	**3.33**	**1.39**	**0.89**	**1.29**	**4.01**	**4.16**	**4.20**	–	–	–	–	–	–
With a speaker prompt															
AudioLM	3.77	3.40	3.75	1.50	1.47	1.54	**3.91**	**4.06**	4.10	0.46	0.48	0.48	**0.96**	0.91	0.86
SoundStorm	**2.99**	**2.43**	**3.36**	**1.10**	**0.81**	**1.24**	3.81	**4.05**	**4.15**	**0.57**	**0.59**	**0.59**	**0.96**	**0.94**	**0.91**

Results on LibriSpeech test-clean

System	SIM-O ↑	WER ↓	FSD ↓	SMOS ↑	CMOS ↑
LibriSpeech *test-clean*					
AR + SoundStorm	0.672	3.267	0.998	$4.20_{\pm 0.17}$	$-0.02_{\pm 0.20}$
MaskGCT	**0.687**	**2.634**	**0.886**	$\mathbf{4.27}_{\pm \mathbf{0.14}}$	$\mathbf{0.10}_{\pm \mathbf{0.16}}$
SeedTTS *test-en*					
AR + SoundStorm	0.683	2.846	0.323	$4.03_{\pm 0.23}$	$-0.05_{\pm 0.22}$
MaskGCT	**0.717**	**2.623**	**0.188**	$\mathbf{4.24}_{\pm \mathbf{0.12}}$	$\mathbf{0.03}_{\pm \mathbf{0.14}}$
SeedTTS *test-zh*					
AR + SoundStorm	0.747	3.865	0.238	$3.78_{\pm 0.23}$	$-0.32_{\pm 0.19}$
MaskGCT	**0.774**	**2.273**	**0.106**	$\mathbf{4.09}_{\pm \mathbf{0.12}}$	$\mathbf{0.05}_{\pm \mathbf{0.17}}$

Mask-GCT

System	SIM-O ↑	WER ↓	FSD ↓	SMOS ↑	CMOS ↑
LibriSpeech *test-clean*					
Ground Truth	0.68	1.94	-	$4.05_{\pm 0.12}$	0.00
VALL-E [2]	0.50	5.90	-	$3.47_{\pm 0.26}$	$-0.52_{\pm 0.22}$
VoiceBox [9]	0.64	2.03	0.762	$3.80_{\pm 0.17}$	$-0.41_{\pm 0.13}$
NaturalSpeech 3 [8]	0.67	**1.94**	0.786	$4.26_{\pm 0.10}$	$\mathbf{0.16}_{\pm 0.14}$
VoiceCraft [5]	0.45	4.68	0.981	$3.52_{\pm 0.21}$	$-0.33_{\pm 0.16}$
XTTS-v2 [52]	0.51	4.20	0.945	$3.02_{\pm 0.22}$	$-0.98_{\pm 0.19}$
MaskGCT	0.687(0.723)	2.634(1.976)	0.886	$4.27_{\pm 0.14}$	$0.10_{\pm 0.16}$
MaskGCT (*gt length*)	**0.697**	2.012	**0.746**	$\mathbf{4.33}_{\pm 0.11}$	$0.13_{\pm 0.13}$
SeedTTS *test-en*					
Ground Truth	0.730	2.143	-	$3.92_{\pm 0.15}$	0.00
CosyVoice [53]	0.643	4.079	0.316	$3.52_{\pm 0.17}$	$-0.41_{\pm 0.18}$
XTTS-v2 [52]	0.463	3.248	0.484	$3.15_{\pm 0.22}$	$-0.86_{\pm 0.19}$
VoiceCraft [5]	0.470	7.556	0.226	$3.18_{\pm 0.20}$	$-1.08_{\pm 0.15}$
MaskGCT	0.717(0.760)	2.623(1.283)	0.188	$\mathbf{4.24}_{\pm 0.12}$	$0.03_{\pm 0.14}$
MaskGCT (*gt length*)	**0.728**	**2.466**	**0.159**	$4.13_{\pm 0.17}$	$\mathbf{0.12}_{\pm 0.15}$
SeedTTS *test-zh*					
Ground Truth	0.750	1.254	-	$3.86_{\pm 0.17}$	0.00
CosyVoice [53]	0.750	4.089	0.276	$3.54_{\pm 0.12}$	$-0.45_{\pm 0.15}$
XTTS-v2 [52]	0.635	2.876	0.413	$2.95_{\pm 0.18}$	$-0.81_{\pm 0.22}$
MaskGCT	0.774(0.805)	2.273(0.843)	0.106	$4.09_{\pm 0.12}$	$0.05_{\pm 0.17}$
MaskGCT (*gt length*)	**0.777**	**2.183**	**0.101**	$\mathbf{4.11}_{\pm 0.12}$	$\mathbf{0.08}_{\pm 0.18}$

FSD - Fréchet Speech Distance computed with wav2vec2.0 features in a manner similar to FID for images

F5-TTS

Model	WER(%)↓	SIM-o↑	CMOS↑	SMOS↑
Seed-TTS *test-en*				
Ground Truth	2.06	0.73	0.00	3.91
Vocoder Resynthesized	2.09	0.70	-	-
CosyVoice	3.39	0.64	0.02	3.64
FireRedTTS	3.82	0.46	-1.46	2.94
MaskGCT	2.623*	0.717*	-	-
Seed-TTS$_{DiT}$	**1.733***	**0.790***	-	-
E2 TTS (32 NFE)	2.19	0.71	0.06	3.81
F5-TTS (16 NFE)	1.89	0.67	0.16	3.79
F5-TTS (32 NFE)	1.83	0.67	**0.31**	**3.89**
Seed-TTS *test-zh*				
Ground Truth	1.26	0.76	0.00	3.72
Vocoder Resynthesized	1.27	0.72	-	-
CosyVoice	3.10	0.75	-0.06	3.54
FireRedTTS	1.51	0.63	-0.49	3.28
MaskGCT	2.273*	0.774*	-	-
Seed-TTS$_{DiT}$	**1.178***	**0.809***	-	-
E2 TTS (32 NFE)	1.97	0.73	-0.04	3.44
F5-TTS (16 NFE)	1.74	0.75	0.02	3.72
F5-TTS (32 NFE)	1.56	0.76	**0.21**	**3.83**

Model	#Param.	#Data	WER(%)↓	SIM-o↑	RTF↓
LibriSpeech *test-clean*					
Ground Truth (*2.2 hours subset*)			2.2	0.754	-
VALL-E 2	-	50K EN	2.44	0.643	0.732
MELLE	-	50K EN	2.10	0.625	0.549
MELLE-*R2*	-	50K EN	2.14	0.608	0.276
Voicebox	330M	60K EN	**1.9**	**0.662**	0.64
DiTTo-TTS	740M	55K EN	2.56	0.627	**0.162**
Ground Truth (*40 samples subset*)			1.94	0.68	-
Voicebox	330M	60K EN	2.03	0.64	0.64
NaturalSpeech 3	500M	60K EN	**1.94**	0.67	0.296
MaskGCT	1048M	100K Multi.	2.634	**0.687**	-
LibriSpeech-PC *test-clean*					
Ground Truth (*1127 samples 2 hrs*)			2.23	0.69	-
Vocoder Resynthesized			2.32	0.66	-
CosyVoice	∼300M	170K Multi.	3.59	0.66	0.92
FireRedTTS	∼580M	248K Multi.	2.69	0.47	0.84
E2 TTS (32 NFE)	333M	100K Multi.	2.95	**0.69**	0.68
F5-TTS (16 NFE)	336M	100K Multi.	2.53	0.66	**0.15**
F5-TTS (32 NFE)	336M	100K Multi.	**2.42**	0.66	0.31

CosyVoice 2, FishSpeech

Model	*test-zh*		*test-en*		*test-hard*	
	CER (%) ↓	SS ↑	WER (%) ↓	SS ↑	WER (%) ↓	SS ↑
Human	1.26	0.755 (0.775)	2.14	0.734 (0.742)	-	-
Vocoder Resyn.	1.27	0.720	2.17	0.700	-	-
Seed-TTS† [33]	1.12	0.796	2.25	0.762	7.59	0.776
FireRedTTS [35]	1.51	0.635 (0.653)	3.82	0.460 (0.526)	17.45	0.621 (0.639)
MaskGCT [18]	2.27	0.774 (0.752)	2.62	0.714 (0.730)	10.27	0.748 (0.720)
E2 TTS (32 NFE)† [31]	1.97	0.730	2.19	0.710	-	-
F5-TTS (32 NFE) [32]	1.56	0.741 (0.794)	1.83	0.647 (0.742)	8.67	0.713 (0.762)
CosyVoice [34]	3.63	0.723 (0.775)	4.29	0.609 (0.699)	11.75	0.709 (0.755)
CosyVoice 2	1.45	0.748 (0.806)	2.57	0.652 (0.736)	6.83	0.724 (0.776)
CosyVoice 2-S	1.45	0.753 (0.812)	2.38	0.654 (0.743)	8.08	0.732 (0.785)

Model Name	WER(%) [2]
Ground Truth	**9.22**
fish-speech	**6.89**
reecho	11.92
F5-TTS	13.98
CosyVoice	22.20

Model Name	Resemblyzer	SpeechBrain[3]
Ground Truth	**0.921**	**0.770**
CosyVoice	0.936	0.813
fish-speech	**0.914**	**0.762**
F5-TTS	0.905	0.787
reecho	0.887	0.636

Evaluation of speaker similarity

Model Name	MOS
Ground Truth	**5.00**
fish-speech	**4.05**
CosyVoice	3.80
F5-TTS	2.90
reecho	3.76

Llasa

Model	test-zh		test-en		test-hard	
	CER ↓	sim-o ↑	WER ↓	sim-o ↑	WER ↓	sim-o ↑
Human	1.26	0.755	2.14	0.734	-	-
Our Codec Resyn.	1.92	0.677	2.91	0.619	-	-
Seed-TTS	1.12	0.796	2.25	0.762	7.59	0.776
FireRedTTS	1.51	0.635	3.82	0.460	17.45	0.621
MaskGCT	2.27	0.774	2.62	0.714	10.27	0.748
E2 TTS (32 NFE)	1.97	0.730	2.19	0.710	-	-
F5-TTS (32 NFE)	1.56	0.741	1.83	0.6147	8.67	0.713
CosyVoice	3.63	0.723	4.29	0.609	11.75	0.709
CosyVoice 2	1.45	0.748	2.57	0.652	6.83	0.724
Train-time Scaling						
Llasa-1B-80k	2.69	0.648 (0.779)	3.71	0.541 (0.685)	17.11	0.618 (0.765)
Llasa-1B-160k	2.22	0.658 (0.783)	3.60	0.563 (0.701)	16.73	0.627 (0.770)
Llasa-1B-250k	1.89	0.669 (0.794)	3.22	0.572 (0.708)	12.13	0.638 (0.779)
Llasa-3B-250k	1.60	0.675 (0.792)	3.14	0.579 (0.708)	13.37	0.652 (0.782)
Llasa-8B-250k	1.59	0.684 (0.798)	2.97	0.574 (0.706)	11.09	0.660 (0.787)
Partial PRM (spk sim)						
Llasa-1B-80k	1.52	0.811 (0.849)	2.30	0.761 (0.798)	16.09	0.759 (0.824)
Llasa-1B-160k	1.29	0.815 (0.851)	2.29	0.774 (0.804)	14.10	0.768 (0.830)
Llasa-1B-250k	1.11	0.818 (0.855)	2.03	0.781 (0.809)	11.30	0.773 (0.833)
Llasa-3B-250k	1.06	0.824 (0.856)	1.89	**0.784** (0.812)	11.22	0.780 (0.836)
Llasa-8B-250k	1.04	**0.827** (0.856)	1.84	0.783 (0.806)	10.59	**0.785** (0.839)
Partial PRM (spk sim)+ORM (WER)						
Llasa-1B-80k	0.53	0.809 (0.840)	1.43	0.761 (0.792)	7.22	0.732 (0.789)
Llasa-1B-160k	0.53	0.812 (0.841)	1.49	0.775 (0.798)	6.35	0.745 (0.799)
Llasa-1B-250k	**0.45**	0.818 (0.845)	1.46	0.782 (0.801)	5.24	0.750 (0.803)
Llasa-3B-250k	0.50	0.823 (0.848)	**1.31**	0.783 (0.803)	5.39	0.759 (0.808)
Llasa-8B-250k	0.47	0.825 (0.848)	1.39	0.783 (0.799)	**4.38**	0.767 (0.812)
Llasa-8B-250k	Chunking: if len(char) $>$ 100 $\rightarrow$ 2 chunks, $>$ 200 $\rightarrow$ 3 chunks, . . .				**3.12**	0.770 (0.791)

Spark-TTS

Model	test-zh		test-en	
	CER↓	SIM↑	WER↓	SIM↑
Multi-Stage or NAR Methods				
Seed-TTS†	**1.12**	**0.796**	2.25	**0.762**
FireRedTTS	1.51	0.635	3.82	0.460
MaskGCT	2.27	0.774	2.62	0.714
E2 TTS (32 NFE)†	1.97	0.730	2.19	0.710
F5-TTS (32 NFE)	1.56	0.741	**1.83**	0.647
CosyVoice	3.63	0.723	4.29	0.609
CosyVoice2	1.45	0.748	2.57	0.652
One-Stage AR Methods				
Llasa-1B-250k	1.89	0.669	3.22	0.572
Llasa-3B-250k	1.60	0.675	3.14	0.579
Llasa-8B-250k	1.59	**0.684**	2.97	0.574
Spark-TTS	**1.20**	0.672	**1.98**	**0.584**

Evaluation on the SEED test sets

MegaTTS-3

Model	#Params	Training Data	SIM-O↑	SIM-R↑	WER↓	CMOS↑	SMOS↑	RTF↓
GT	-	-	0.68	-	1.94%	+0.12	3.92	-
VALL-E 2*	0.4B	LibriHeavy	0.64	0.68	2.44%	-	-	-
VoiceBox†	0.4B	Collected (60kh)	0.64	0.67	2.03%	-0.20	3.81	0.340
DiTTo-TTS*	0.7B	Collected (55kh)	0.62	0.65	2.56%	-	-	-
NaturalSpeech 3†	0.5B	LibriLight	0.67	0.76	**1.81%**	-0.10	3.95	0.296
CosyVoice	0.4B	Collected (172kh)	0.62	-	2.24%	-0.18	3.93	1.375
MaskGCT	1.0B	Emilia (100kh)	0.69	-	2.63%	-	-	-
F5-TTS	0.3B	Emilia (100kh)	0.66	-	1.96%	-0.12	3.96	0.307
MegaTTS 3	0.3B	LibriLight	**0.71**	**0.78**	**1.82%**	**0.00**	**3.98**	0.188
MegaTTS 3-accelerated	0.3B	LibriLight	0.70	**0.78**	1.86%	-0.03	3.96	**0.124**

Zero-shot TTS results on the LibriSpeech test-clean set

Models	Tokens/s	Latent Layer	Type	PESQ↑	STOI↑	ViSQOL↑	MCD↓	UTMOS↑
Encodec	600	8	Discrete	3.16	0.94	4.31	1.63	3.07
DAC	450	9	Discrete	**4.13**	**0.97**	4.68	1.05	4.01
WavTokenizer	75	1	Discrete	2.55	0.88	3.83	1.99	4.07
X-codec2	50	1	Discrete	3.03	0.91	4.12	1.72	**4.13**
WaveVAE	25	1	Continuous	3.84	0.96	**4.71**	**1.03**	4.10

Comparison of the reconstruction quality

VibeVoice

Model	Frame Rate	test-zh		test-en	
		CER(%) ↓	SIM ↑	WER(%) ↓	SIM ↑
MaskGCT [WZL+24]	50	2.27	0.774	2.62	0.714
Seed-TTS [ACC+24b]	-	1.12	0.796	2.25	0.762
FireRedTTS [GLS+24]	25	1.51	0.635	3.82	0.460
CosyVoice 2 [DWC+24b]	25	1.45	0.748	2.57	0.652
Spark TTS [WJM+25]	50	1.20	0.672	1.98	0.584
VIBEVOICE-1.5B	7.5	1.16	0.744	3.04	0.689

Results on the SEED test sets

Tokenizer	N_q	Token Rate	test-clean			test-other		
			PESQ	STOI	UTMOS	PESQ	STOI	UTMOS
Ground-Truth	-	-	-	-	4.056	-	-	3.483
Encodec [DCSA22]	8	600	2.72	**0.939**	3.04	2.682	**0.924**	2.657
DAC [KSL+23]	4	400	2.738	0.928	3.433	2.595	0.908	2.945
Encodec [DCSA22]	4	300	2.052	0.901	2.307	2.052	0.884	2.088
SpeechTokenizer [ZZL+23]	4	300	1.931	0.878	3.563	1.737	0.837	3.018
DAC [KSL+23]	1	100	1.246	0.771	1.494	1.245	0.751	1.499
WavTokenizer [JJW+25]	1	75	2.373	0.914	4.049	2.261	0.891	3.431
WavTokenizer [JJW+25]	1	40	1.703	0.862	3.602	1.662	0.834	3.055
Ours (Acoustic)	1	7.5	**3.068**	0.828	**4.181**	**2.848**	0.823	**3.724**

Comparison of the reconstruction quality

Questions?

tg: @elagutin