Hierarchical Semantic-Acoustic Modeling via Semi-Discrete Residual Representations for Expressive End-to-End Speech Synthesis

Yixuan Zhou[1], Guoyang Zeng[2], Xin Liu[2], Xiang Li[1], Renjie Yu[1], Ziyang Wang[2], Runchuan Ye[1], Weiyue Sun[2], Jiancheng Gui[2], Kehan Li[1], Zhiyong Wu[1,], Zhiyuan Liu[3]*

[1] Shenzhen International Graduate School, Tsinghua University [2] ModelBest Inc

[3] Department of Computer Science and Technology, Tsinghua University

yx-zhou23@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

GitHub Samples Demo

ICLR VoxCPM THUHCSI OpenBMB ModelBest

Introduction

Problem: Expressivity-Stability Trade-off in TTS

- **Discrete tokens** (VALL-E, SparkTTS, etc.): stable generation, but **quantization ceiling** discards acoustic detail
  - Direct FSQ/VQ at speech signals for LM target → **codebook explosion** when needing more information
- **Continuous tokens** (DiTAR, VibeVoice, etc.): retain acoustic richness, but **task entanglement** → error accumulation
- **Multi-stage pipelines** (CosyVoice, IndexTTS2, etc.): **semantic-acoustic divide** (LM&diffusion) prevents end-to-end optimization

Contributions

- Propose **end-to-end hierarchical architecture** with differentiable **semi-discrete bottleneck** (FSQ)
  → resolves expressivity-stability trade-off, no external discrete speech tokenizer
- Introduce **residual learning** (TSLM + RALM) → functional separation without architectural fragmentation
- Achieve **SOTA zero-shot TTS** among open-source systems (0.5B, 1M+ hours)
- Provide extensive ablation studies validating semi-discrete residual representations as crucial for robust, expressive synthesis
  → Visualization & Probing validate **emergent specialization**: TSLM ~ semantics & prosody, RALM ~ acoustic details

Methodology

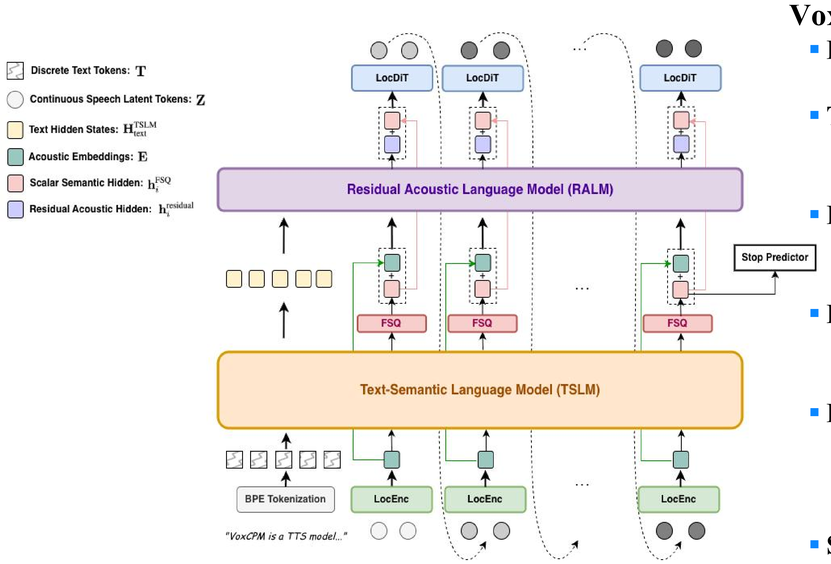


Fig.1: Overall Architecture

VoxCPM Architecture

- **LocEnc**
  - Bidirectional Transformer: get patch-level acoustic emb
- **TSLM (Text-Semantic LM)**
  - Global causal backbone for semantic-prosodic planning
  - Initialized from pre-trained text LM (MiniCPM4)
- **FSQ Bottleneck**
  - **semi-discrete**: larger dimension (256-d) than normal FSQ
  - Induces natural TSLM/RALM specialization
- **RALM (Residual Acoustic LM)**
  - Recovers fine-grained acoustic details dropped by FSQ
  - Residual input fusion: re-injects acoustic emb
- **LocDiT**
  - Bidirectional Transformer: patch-level AudioVAE latents reconstruction with flow matching
  - conditioned by TSLM+RALM hiddens, with CFG strategy
- **Stop Predictor**
  - Simple linear layer for stop prediction at TSLM hidden
- **AudioVAE**
  - Causal CNN, transform waveform to 25 Hz, 64-d latents
  - Mel reconstruction loss + Discriminator loss + KL loss

Training:

- VAE pre-trained separately; than VoxCPM backbone
- **Simple objective**: flow-matching loss & stop CE loss
- **WSD learning rate schedule:** for better performance
- Cost: 40 H100 GPUs, 1 week (1M+ hours speech data)

The Key Insight: Quantization as Inductive Bias, Not Prediction Target

- FSQ sits **between** hierarchical modules as a differentiable bottleneck within a unified model
- Quantization is **NOT** the prediction target — it constrains the hidden state space
- Gradients flow through FSQ via **straight-through estimation** → fully end-to-end

Experiments

Main Results (Seed-TTS-Eval)

- **SOTA** in both intelligibility and speaker similarity among open-source TTS systems

Model	Params	Data/hrs	EN		ZH		Hard	
			WER ↓	SIM ↑	CER ↓	SIM ↑	CER ↓	SIM ↑
F5-TTS	0.3B	100K	2.00	67.0	1.53	76.0	8.67	71.3
MaskGCT	1B	100K	2.62	71.7	2.27	**77.4**	-	-
CosyVoice	0.3B	170K	4.29	60.9	3.63	72.3	11.75	70.9
CosyVoice2	0.5B	170K	3.09	65.9	1.38	75.7	**6.83**	72.4
SparkTTS	0.5B	100K	3.14	57.3	1.54	66.0	-	-
FireRedTTS	0.5B	248K	3.82	46.0	1.51	63.5	17.45	62.1
FireRedTTS-2	-	1.4M	1.95	66.5	1.14	73.6	-	-
Qwen2.5-Omni	7B	-	2.72	63.2	1.70	75.2	7.97	74.7
OpenAudio-s1-mini	0.5B	2M	1.94	55.0	1.18	68.5	23.37	64.3
IndexTTS 2	1.5B	55K	2.23	70.6	1.03	76.5	7.12	**75.5**
VibeVoice	1.5B	-	3.04	68.9	1.16	74.4	-	-
HiggsAudio-v2	3B	10M	2.44	67.7	1.50	74.0	55.07	65.6
VoxCPM-Emilia	0.5B	100K	2.34	68.1	1.11	74.0	12.46	69.8
VoxCPM	0.5B	1.8M	**1.85**	**72.9**	**0.93**	77.2	8.87	73.0

Subjective Evaluation

- Good naturalness in ZH/EN
- Best speaker similarity
- Scaling data is important

Model	ZH		EN	
	N-MOS	S-MOS	N-MOS	S-MOS
MaskGCT	3.20 ± 0.11	3.77 ± 0.11	3.84 ± 0.11	4.00 ± 0.10
CosyVoice 2	3.38 ± 0.12	4.01 ± 0.10	4.14 ± 0.09	3.97 ± 0.10
IndexTTS 2	4.25 ± 0.09	4.05 ± 0.09	4.03 ± 0.10	4.16 ± 0.09
VoxCPM-Emilia	3.79 ± 0.12	3.99 ± 0.11	3.91 ± 0.10	4.10 ± 0.09
VoxCPM	4.10 ± 0.10	4.11 ± 0.10	4.11 ± 0.09	4.18 ± 0.09

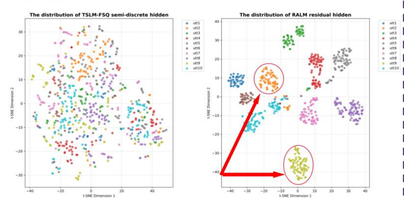
Ablation: Architecture Components

- FSQ dimension
- TSLM initialization
- hierarchy design
- Residual input
- LocDiT condition

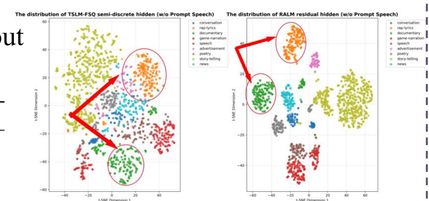
Model Setting	EN		ZH		ZH-hard case	
	WER ↓	SIM ↑	CER ↓	SIM ↑	CER ↓	SIM ↑
default setting (w/ FSQ: d256s9)	2.98	62.6	1.77	70.4	18.19	64.9
w/ FSQ: d4s9	5.18	59.3	4.05	68.0	19.55	62.3
w/ FSQ: d128s9	3.43	62.2	1.67	70.7	16.76	65.7
w/ FSQ: d1024s9	3.07	62.0	2.38	69.8	20.38	64.7
w/o FSQ: d1024s∞	3.67	62.1	2.30	69.6	24.92	63.5
Hierarchical, w/o LM init. in TSLM	5.24	63.4	2.41	70.9	24.66	65.6
w/o RALM: TSLM (24 layers, LM init.) → LocDiT	4.34	61.8	3.05	69.4	25.00	63.8
w/o RALM: TSLM (30 layers, random init.) → LocDiT	5.35	62.6	3.46	69.8	30.40	63.9
w/o RALM: TSLM (30 layers, partial LM init.)→ LocDiT	4.12	62.0	3.07	69.6	26.20	63.1
w/o $E_{<i}$ in RALM: TSLM → ALM → LocDiT	4.91	60.9	4.94	68.1	27.17	61.7
w/o $h^{residual}$ in condition: TSLM → FSQ → LocDiT	3.86	58.3	3.05	67.6	23.65	61.7

Analysis: implicit decoupling of semantic & acoustic

- **t-SNE Visualization**
  - TSLM ~ weak correlation with speaker, strongly correlated with text style & prosody
  - RALM ~ strong correlation with speaker, and infers suitable timbre from text semantics
- **Probing Result (SUPERB)**
  - TSLM ~ best at linguistic content
  - RALM ~ best at speaker identity
  - FSQ bottleneck induces **division** of labor without explicit supervision



(voice cloning, w/ prompt)



(text to speech, w/o prompt)

Hidden State Location	PR (PER) ↓	ASR (WER) ↓	ASV (EER) ↓
LocEnc output	59.12	65.79	15.38
TSLM last hidden (Pre-FSQ)	**45.60**	**60.43**	18.70
FSQ output	50.90	62.37	19.25
RALM last hidden	53.49	64.85	**13.24**

Layer-wise probing results on internal hidden states of VoxCPM