腾讯游戏知几语音合成大模型推理加速方案

李正兴 / 腾讯高级工程师

DataFunSummit # 2024

- 背景介绍
- 语音合成模型结构分析
- 语音合成模型推理思路
- 未来展望

01

背景介绍

背景-产品展示

01、王者荣耀小妲己“游戏知识问答”

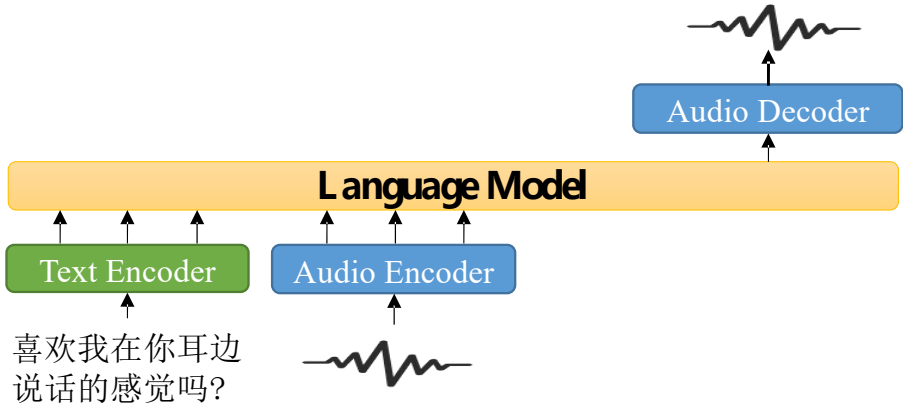
02、和平第五人的『AI语音助手』

03、天涯明月刀『绝智阿暖』智能NPC

背景-产品展示

Tencent 腾讯 互动娱乐 INTERACTIVE ENTERTAINMENT

- TTS：更自然、韵律丰富、更实时
  - 采用LM方案 -- 自研***知音语音大模型***
  - ***10s*** 音频完成声音复刻
  - 通过加速优化，实时率 ***~0.085***



	范闲	老头	云悠悠	英语男	英语女
原音					
CFer你好呀！喜欢姐姐的AK四七吗？不喜欢的话还有M四A一和AN九四哦.姐姐的ASMR你受得了吗？I love you my sweetheart~					
你在开什么玩笑？我才不会上当呢。					

Tencent 腾讯 互动娱乐 INTERACTIVE ENTERTAINMENT

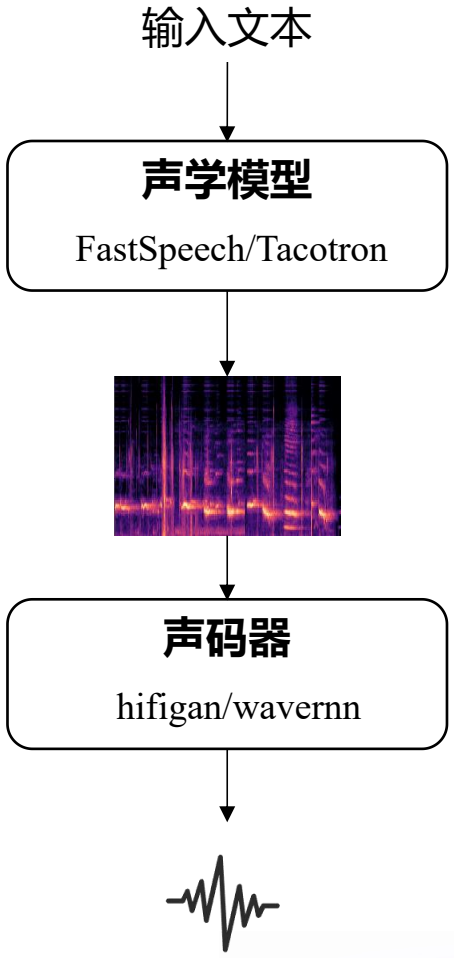
02

模型结构选型与分析

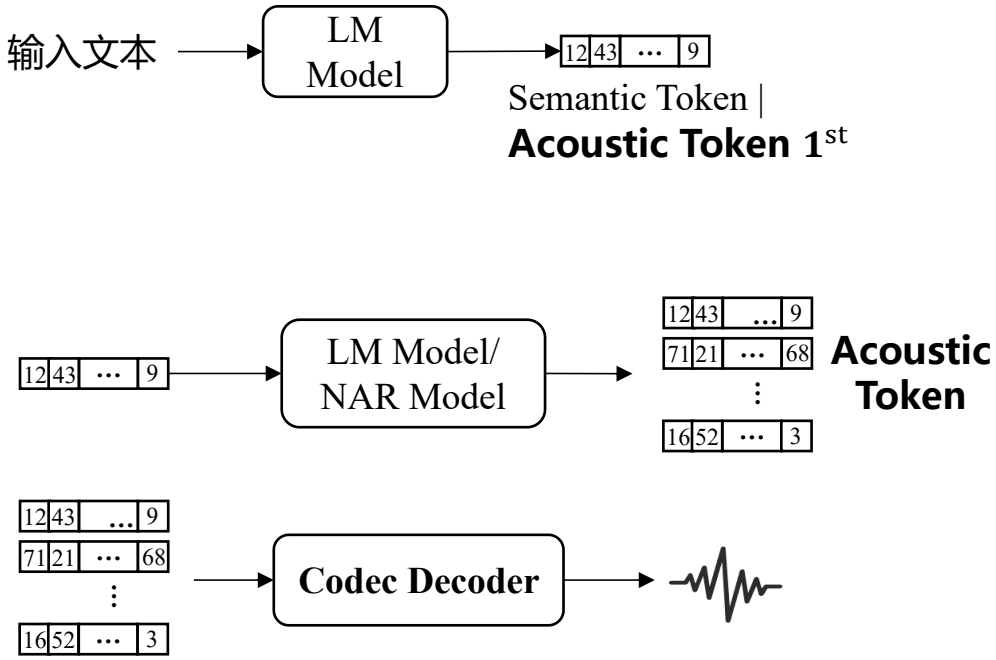
DataFun.

语音合成大模型结构

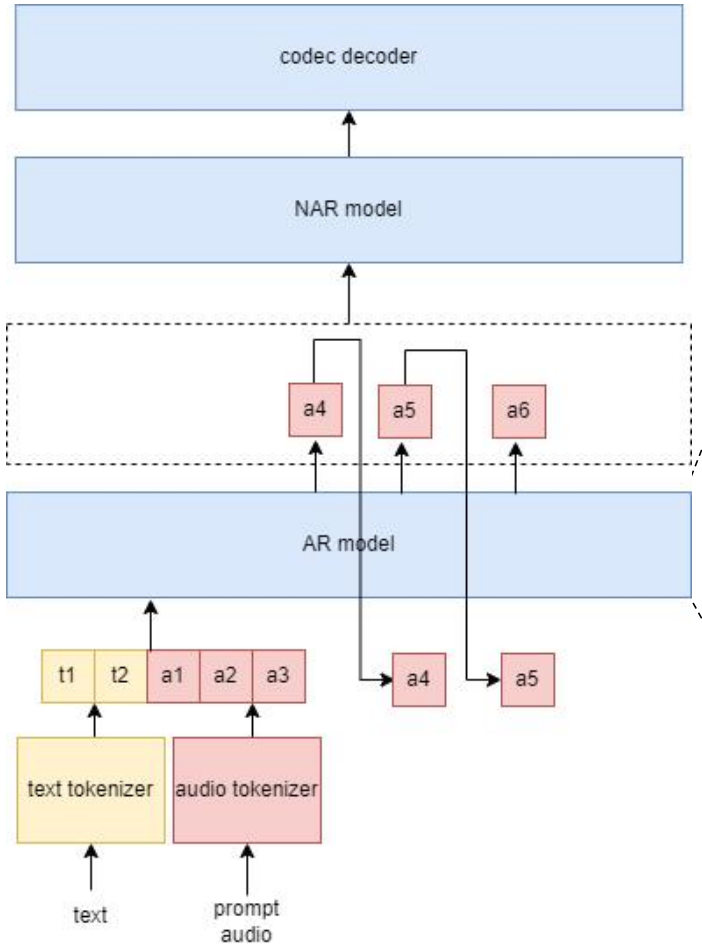
传统方案

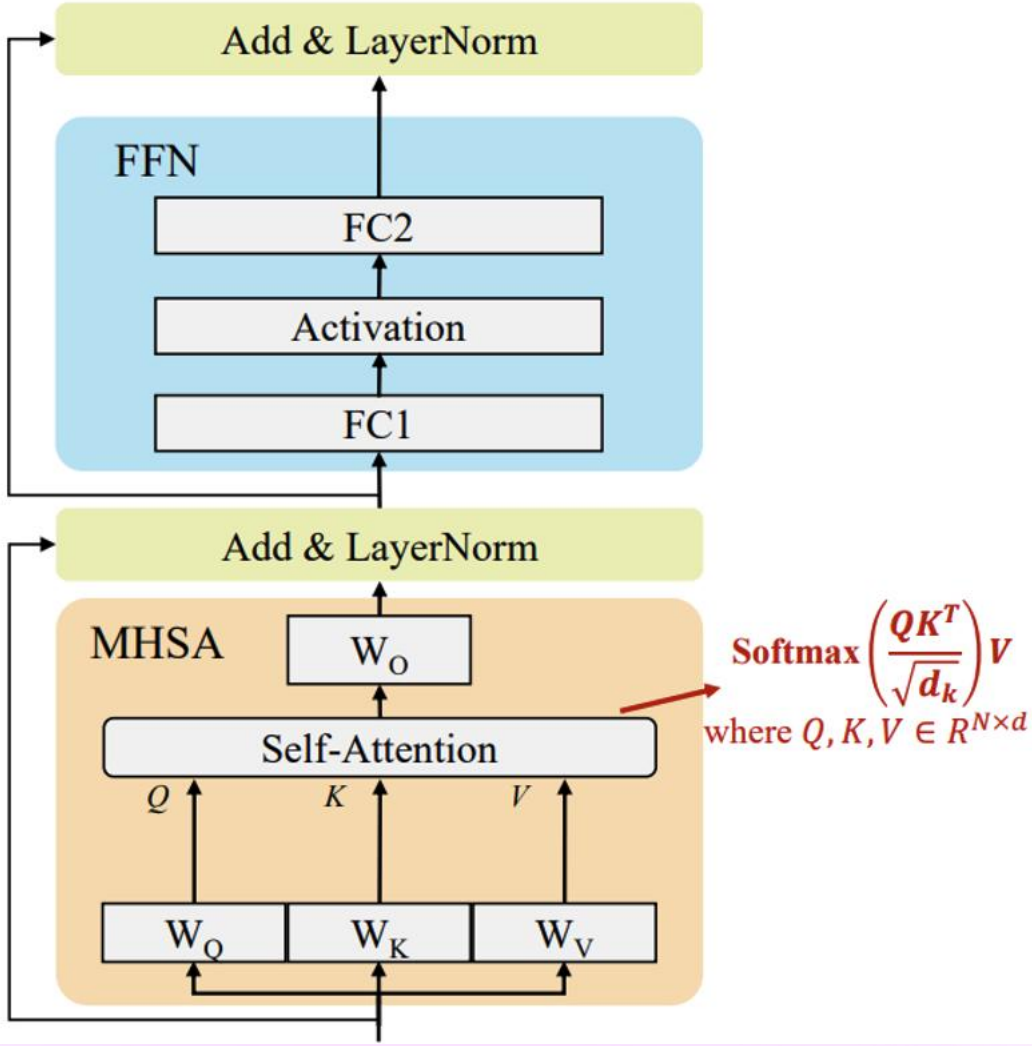


基于语言模型的新方案



语音合成大模型结构





面临的挑战：

1. 高并发场景

2. 实时率问题

03

模型推理加速方案

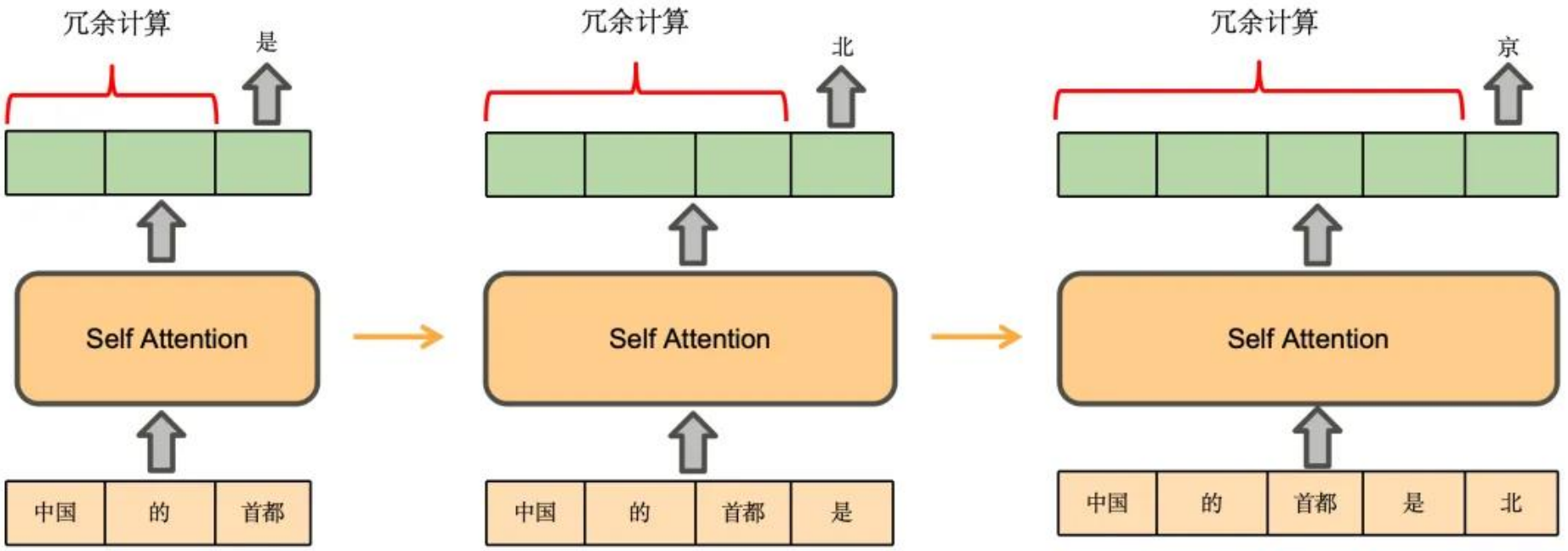
推理加速方案-借鉴与选择

是否能将NLP领域的LLM 推理加速方法应用到语音合成大模型上？

- kv cache
- flash attention
- page attention
- 投机采样
- prefix kv cache
- flash decode
- Int4/int8 量化
- …….

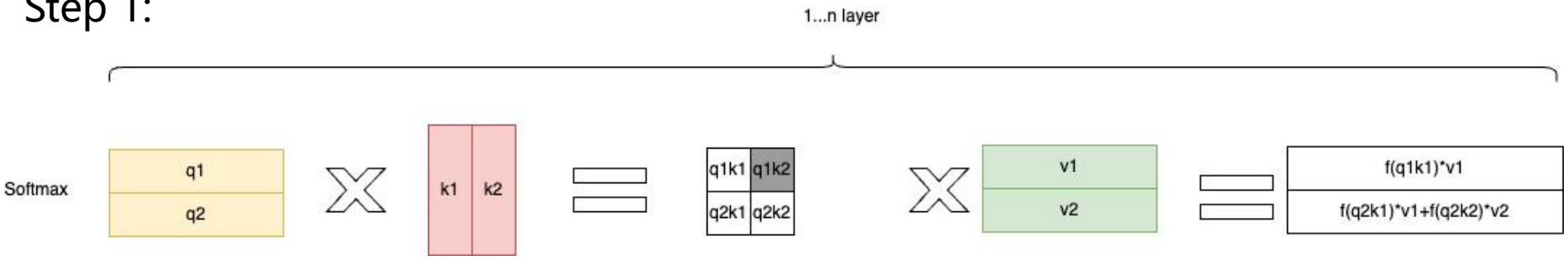
推理加速方案-kv cache

LLM 中的kv cache：

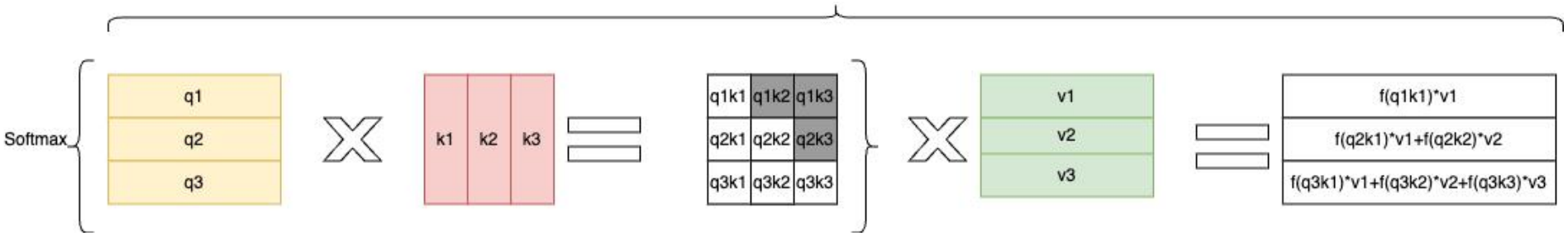


推理加速方案-kv cache

Step 1:



Step 2，without kv cache:



Step 2，with kv cache: