**技术FISH-SPEECH**

娇兰数字人工厂项目组

**2024-1-3**

1. 技术需求

在当前的数字时代，声音定制和优化变得越来越重要。无论是在媒体制作、虚拟助手、在线教育还是在游戏和应用程序开发中，定制化的声音解决方案可以极大地提升用户体验和参与度。我们需要一个性化、动态适应性强并且用户友好的声音定制。主要追求的是：高质量声音合成，个性化选项，情感表达能力。主要核心点为能够产生清晰、自然的人声，允许用户根据个人喜好调整音调、速度、音量和语调，使声音能夠表达不同的情感，如快乐、悲伤、兴奋等。

1. FISH-SPEECH技术

作者：Fish Audio

该作者另一项目为：Bert-VITS2

Free speech,总的来说有两个模块组成：

1. VQGAN：将梅尔普压缩为语义token,再重新解码为音频的模块啊,这一块更接近于一个HIFIGAN的结构。

梅尔谱：基于H\_ling和H\_spk（H\_spk是可训练的矢量），目标梅尔谱（mel-spec）的低频部分通过韵律编码网络获得隐藏韵律矢量（LPV）。最后，H\_ling、H\_spk和隐藏韵律矢量拼接在一起，送入后续的长度规整模块和解码网络获得预测的梅尔谱。

1. 基于LLAMA做的自回归模型

专门为对话领域微调的模型

目前更像一个测试版的模型，主要还是会附带一些电音的问题

需要把文素和文本推入到模型中，然后让模型生成一个量化的语义token，再通过Stage1中训练的decoder把它们进行解码。对于推理来说,我们只需要2GB的显存,对于微调只需要16G或者24G,300~500,token每秒的生成速度,也就是说,单卡一秒钟可以生成10~20秒的音频,这个基本上是可以满足大部分的需求。window下不支持训练,只能推理。GPU要求GPU内存: 2GB (用于推理), 16GB (用于微调) ，linux下可以使用全部功能。

官网地址：<https://speech.fish.audio/>

仓库地址：https://github.com/fishaudio/fish-speech

1. 场景整理

云上部署:https://www.autodl.com/market/list

本地4090: windows

本地4090：linux

1. 商用模型，火山引擎

<https://www.volcengine.com/product/voicecloning>

仅需在开放环境下录制10S的音频即可快速复刻

高度还原真人音色特点，说话风格，口音和声学环境