快速复刻SOP

1. 语料要求

质量要求

- 原声只包含一个说话人
- 无插话或少量语气词插话
- 尽可能贴近人设，语气自然，符合场景
- 语音清晰可懂，无背景音或尽可能少背景音
- 音质稳定，音调不过尖锐
- 完整且连续的一整段内容，如果原始复刻素材句子不完整，有截断，会导致的电平跳变、有噼啪声
- 【prompt复刻】prompt音频末尾需要有500ms以上的空白音，如果末尾没有留白，会把prompt末尾补全，然后生成正常的音频，补全的部分就是噪声。
- 尽量不要只读“一二三四五六七”这样的文案，其中不包含平舌音卷舌音长鼻音后鼻音，模型可能会随机
- speech-2.5增强了位置编码，会还原节奏韵律和口癖，在用2.5preview的时候，需要考虑一下几段素材之间留空长度，以及是不是保留“嗯啊哦唔”等语气词
- 不建议用语种混合的语料，目标语言是什么，就用什么语言的原始语料进行复刻；如果目标语言是中英混杂，建议用纯中文或者纯英文进行复刻

量级及格式要求

- **15s左右间即可，几段几秒不重要，希望截取的是本身自然的一句话**
- **快速复刻只是提取音频里的音调节奏，素材太多了反而会平**
- 单声道，快速复刻的策略是左右声道取平均，如果是单一声道，素材响度就会在正常值之外，整个表现都会变差
- 建议最好是wav格式，44100hz采样率，16bits位深

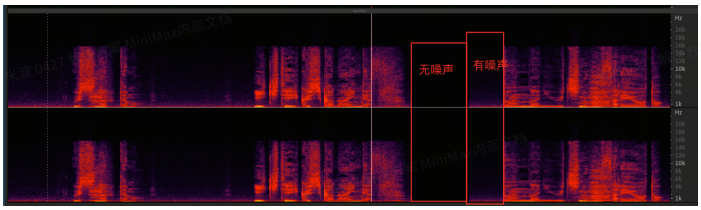
语音录制注意事项

- 找尽量安静的空间录音（不可以是楼道，卫生间这类大混响空间）
- 苹果设备可使用“语音备忘录” 进行录制（可参考官方教程：https://support.apple.com/zh-sg/guide/iphone/iph4d2a39a3b/ios）

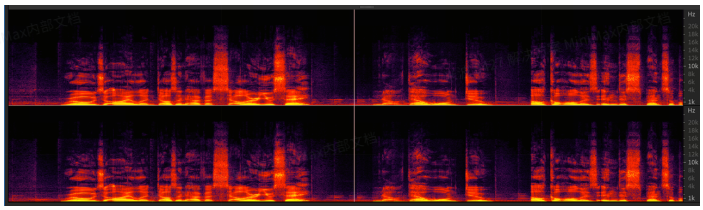
2. 原声预排查

1. **噪声批量过一遍 Audition，需结合频谱图，「听」和「看」一起**
2. **识别噪声的处理**

噪声 vs 非噪声



音乐： 会带下划线



3. 快速复刻小Tips

3.1 复刻素材并非越多越好

快速克隆只是提取音色信息，并不是让大模型“训练”，所以不是给的越多效果越好

不恰当的增加素材，提取出的音色就像是一盒彩色的橡皮泥被混合成棕灰色一样

复刻素材中有九种鲜明的情绪，每一种用来单独复刻都可以完美还原素材中的情绪

但九种情绪素材混合复刻的结果，是情绪平均的，没有语调起伏变化的音色

3.2 除非必要，避免选择混响等声学信息过于突出的素材

我们的模型可以很好的还原混响等声学信息，但其生成同样带有随机性，并且需要连续性才能听起来自然

这意味着如果你需要后期剪辑音频，不连续的空间信息会给你带来麻烦，混响断裂会带来明显不自然

3.3 如果需要更加夸张的相似度，可以选择带prompt的快速复刻

快速复刻可以让模型提取出一个人的声音特征，并根据文本内容展示不同的风格，对于文本更好的响应意味着更高的自然度与更好的表现力

而带prompt的快速复刻则是在prompt音频基础上进行“续写”，情绪/语速/语调等与prompt的相似度更高，但对于文本的响应可能不如快速复刻自然