拟公示算法机制机理内容

算法名称	喜马拉雅语音大模型算法
算法基本原理	喜马拉雅语音大模型算法是一种基于 Transformer 架构的多层模型，使用自有版权的大规模高质量的文本和语音数据进行训练，从而具备对任意文本的音频生成能力和对输入提示音频的音色韵律复刻能力。算法同时会对输入的文本和提示音频进行安全过滤，最终得到安全、准确、高品质的生成音频内容。
算法运行机制	喜马拉雅语音大模型算法，使用特定的授权录音作为提示音频，将输入的文本内容转换成目标说话人的输出语音。喜马拉雅语音大模型主要由 3 部分组成：文本 Token 编码器、音频生成大模型、音频 Token 解码器。 1. 文本 Token 编码器：由两个字典组成，一个是汉字到拼音的字典，另一个是拼音到整数序号的字典，将输入的待合成文本转换为文本 Token。输入数据为待合成文本，例如“今天天气很好。”，输出为离散的整数 token，例如: [10, 25, …, 61]，输出格式为整型数组，输出大小为 1024 个整数以内。 2. 音频生成大语言模型：采用基于 Transformer 结构的机器学习模型，将目标音色音频片断编码为离音频 Token；同时接收第一步文本 Token 和目标音色音频片断，生成待合成文本对应的音频 Token。第一个输入数据为文本 Token 编码器输出的整数 token,例如: [10, 25, …, 61]；第二个输入数据为目标音色音频片断，可以是各类音频格式如“wav”、“mp3”、 “m4a”等，规整为 1024 维度浮点数向量，例如: [0.12, 0.17, …, 0.45]。输出数据为待合成文本对应的音频 Token，例如[132, 190, …, 275]，输出

	格式为整型数组，输出大小为 4096 个整数以内。 3. 音频 Token 解码器：采用基于 GAN+VAE 的机器学习模型，将第 2 步的音频 Token 转换为合成音频。输入数据为第二步音频生成大语言模型输出的音频 Token，例如[132, 190, …, 275]。输出数据为合成语音二进制数据，可以保存为音频文件例如“output.wav”。
算法应用场景	用于喜马拉雅 APP 电子书频道的内容生产
算法目的意图	提升音频生产效率。将语音大模型生成技术应用于有声书、新闻等有声化场景，满足用户实时听书、听新闻的需求。
算法公示情况（选填）	