拟公示算法机制机理内容

算法名称	喜马拉雅语音合成算法
算法基本原理	语音合成技术涉及了语言学知识以及深度学习算法。通过查字典的方式得到文本的读音，之后通过基于深度学习的方法，将读音信息转换成语音信息。生成音频后，会对静音等部分做人工的物理剪辑。暂无其他额外的信号处理。

算法运行机制	语音合成技术(TTS，Text-to-Speech)，可以将文本内容转换成语音。首先“从文本生成拼音序列”模块，将输入的文本，转换成拼音序列，例如待合成“今天天气很好”转换为”jin1tian1tian1qi4hen2hao3”。接着”拼音序列生成频谱信息”模块，将拼音序列转换成频谱信息。最后“频谱信息生成语音信号”模块将频谱信息转换成语音信号。通过如上三个系统，共同完成语音合成的任务。
	下面分别就“发音字典”、“声学模型”、“声码器”三方面进行说明:
	1．发音字典，汉字到拼音的映射规则。发音字典是完全的规则处理。字典示例:
	{
	“今”：“jin1”，
	“天”：“tian1”，
	...
	}
	该字典由人工整理和标注完成，字典包含了 GB18030 中的 27533 个中文汉字符号，及其对应的拼音。
	在实际语音合成时，待合成的文本会逐字查字典，得到拼音序列。

	2. 声学模型，Encoder-Decoder 结构，通过听力测试来评价效果。喜马拉雅语音合成系统中的声学模型结构采用 DurIAN 结构: https://arxiv.org/pdf/1909.01700.pdf， 使用目标音色的 10 小时语音数据进行训练。 声学模型的训练分为特征提取和训练两个步骤。特征提取会使用 librosa 从音频中提出频谱特征，从拼音标注中提取出拼音序列。模型训练会将拼音序列输入 DurIAN 的 Encoder, 并在 DurIAN 的 decoder 输出预测的频谱特征。预测的频谱与真实的频谱特征计划 MSE loss, 直到 Loss 收敛，完成训练。 声学模型的推理过程为：输入发音字典模块输出的拼音序列，经已经训练好的 DurIAN 模型，推理得到频谱特征。频谱特征会由后文将提到的声码器转为音频信号，即合成的语音。 3. 声码器模型，GAN 结构，通过听力测试来评价效果。喜马拉雅语音合成系统使用 HiFiGAN: https://arxiv.org/abs/2010.05646，使用目标音色的 10 小时语音数据进行训练。 声码器的训练分为特征提取和训练两个步骤。特征提取会使用 librosa 从音频中提出频谱特征。模型训练将频谱特征输入 HiFiGAN 的生成器，预测生成音频信号。训练 Loss 参考 HiFiGAN 的 loss, 采用开源实现：https://github.com/jik876/hifi-gan 声码器的推理会使用已经训练好的 HiFiGAN，输入为声学模型预测的频谱特征，输出为模型预测的音频信号。
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。将语音合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）	