语音克隆的基本步骤

- 语音克隆的基本步骤可以大致分成 **语音到文本 文本到语音** 两阶段。

语音到文本（STT）

语音克隆需要从目标说话者的语音中提取文字及其特征，但这里的语音到文本技术并不是简单的语音识别，后者只是将语音转化为纯文本，忽略了说话人的个性化特征，而前者则是集声纹提取、语音特征提取、说话人适配等技术于一体的能够完整提取语音的内容、音色、频率等特征并将其适配到现有语音合成模型的综合技术。

- ## 声纹提取

  - **目标**：从输入语音中提取说话者的个性化声音特征（如音色、语速、语调等），并将其表示为一个紧凑的向量（声纹嵌入，Speaker Embedding）。
  - **常用技术**：
    - 深度学习模型：d-vector和x-vector技术等。
    - 预训练的说话人编码模型：Deep Speaker、ECAPA-TDNN等。
  - **输出**：一个代表目标声音特征的嵌入向量。

- ## 语音特征提取

  - **目标** ：从原始语音信号中提取频谱信息，如**梅尔频谱图**或其他语音特征，获取语音的时间和频率结构
  - **常用工具**：短时傅里叶变换（STFT）提取语谱图

- ## 说话人适配

  - **目标**：将输入的语音特征适配到现有的语音合成模型，使其能够学习目标说话人的特性
  - **方法**：
    - 使用少量目标语音进行模型微调
    - 零样本语音克隆，直接使用嵌入向量生成目标语音

文本到语音（TTS）

这一阶段就是传统的语音合成阶段，文本即为第一阶段获取的文本

- ## 常用方法：

  - **端到端的语音合成**：如Tacotron2、FastSpeech、VITS

- **神经声码器**：如WaveNet、HiFi-GAN，将频谱图转化为语音波形
- **使用模型加速技术**（如模型量化、ONNX加速）提升语音生成速度，支持实时语音克隆应用

- ## 关键技术

  [@WeiYuYinHeChengJiShuZongShuJiYanJiuXianZhuang2020]

  - **共振峰合成**：不同人的语音有不同的共振峰模式，可以从中提取参数经过变换合成语音
  - **波形拼接**：
    - 前期准备：
      - 将语音单元进行切分
      - 将切分好的单元构建成语音库
    - 合成阶段：
      - 选出合成单元，从语音库中提取
      - 按要求进行变换
      - 重叠相加、输出合成语音
  - **谐波加噪声模型（HNM）**：将信号按频率高低分为谐波+噪声两种成分，是合成中的声音更加自然
  - **神经网络及深度神经网络模型（DNN）**：先做非监督学习，学习到的内容作为监督学习的初值进行训练。

详细阐述

声纹提取

[@ZhuHaoBing]
[@FangAnDongFuZaBeiJingXiaShengWenTeZhengTiQuYuShiBie2014a]

- ## 定义

  - 声纹，即说话人语音频谱的信息图，不同的人音色、音调不尽相同，因此具有不同的声纹。
  - 声纹提取，是从语音信号中提取具有区分个体身份特征的声纹特征，可以用于语音识别、语音克隆等领域。

- ## 声纹识别系统概述

  - **技术实现【3】**
    - 自动说话人确认技术（ASV）（一对一）
    - 自动说话人辨认技术（ASI）（一对多）

- **声音来源【4】**
  - 文本提示型：被鉴别人需要根据给定的文字进行发音判别
  - 文本相关型：系统录制被鉴别人的规定文本内容的声音，发出相关内容声音即可识别
  - 文本无关型：不规定说话人发音内容，直接识别声音
- **目标对象【7.8】**
  - 闭集识别：已有范围。
  - 开集识别：可能不在集合内，需要重新记录和训练。

关键技术

分为**语音特征参数提取技术**和**模式匹配识别判断技术**，其中前者主要体现在**语音频谱参数**、**线性预测参数**、**小波特征参数**等方面的参数提取；后者主要有**矢量化模型**、**随机模型**、**神经网络模型**等模型。

语音特征参数提取技术

这里主要介绍线性预测参数中MFCC、PLP的方法

- **梅尔频率倒谱系数（MFCC）**
  - **原理**：通过模拟人耳的听觉感知，将语音信号转化为人耳更易感知的低维特征
  - **主要步骤**：分帧加窗——快速傅里叶变换（FFT）——功率谱计算——映射到梅尔频率，划分滤波器——对数运算模拟人耳——离散余弦变换（DCT），倒谱压缩
  - **特点**：模拟人耳感知，降维并保留关键信息，有良好的鲁棒性
  - **缺点**：对噪声比较敏感，无法捕捉时序信息（此时需要计算一阶和二阶差分以捕捉动态变化）
  - **在声纹识别中的应用**：捕捉语音的个体特征，以区分不同的说话者
- **感知线性预测（PLP）**
  - **原理**：与MFCC类似，模拟人耳听觉的原理，对语音信号进行一系列感知处理，能够提取更加鲁棒性的语音特征
  - **主要步骤**：分帧加窗——快速傅里叶变换（FFT）——功率谱计算——映射到Bark频率——校正、根号运算模拟人耳——**去共振处理**——线性预测——倒谱压缩
  - **特点**：模拟人耳感知，对噪声和说话人的变化具有一定的鲁棒性
  - **缺点**：对高频信息有一定缺失，步骤较为复杂
  - **应用**：与MFCC类似，但在嘈杂环境中表现更佳

模式匹配识别判断技术

这里主要介绍 i-vector 和 **x-vector** 的相关技术*

- **i-vector**

        - **原理**：基于高斯混合模型和因子分析，将语音的高维特征映射到一个低维潜在的特征空间中，同时捕捉说话人特性，将变长语音输入转化为固定长度的特征向量
        - **主要步骤**：
            - **特征提取**：提取短时特征（如MFCC）
            - **UBM（通用背景模型）训练**：使用大量数据训练一个**高斯混合模型（GMM）**，称为UBM
            - **超向量生成**：根据UBM计算形成高位的supervector
            - **因子分析**：通过分解超向量，映射出低维因子向量，即i-vector
        - **优点**：计算高效、适应性强（不同长度的语音输入）
        - **缺点**：噪声敏感
    - **x-vector**
        - **原理**：基于深度神经网络（DNN），从语音中提取固定长度的嵌入向量（输出），学习说话人相关的判别特征
        - **主要步骤**：
            - **特征提取**：提取短时特征（如MFCC）
            - **DNN建模**：使用一个 DNN 模型，其中前几层负责提取局部时间特征，后几层聚合全局信息
            - **统计池化**：对网络进行全局统计，得到固定长度的表示
            - **提取向量**：从神经网络的嵌入层提取x-vector
        - **优点**：判别能力强、鲁棒性好，端到端训练
        - **缺点**：计算复杂度高、数据需求量大

- ## 技术难点

    - 用较短的语音进行模型训练
    - 有效区分真正的声音和模仿的声音
    - 消除或减弱声音变化带来的影响
    - 消除信道差异和背景噪音带来的影响

语音特征提取

将原始语音信号转换为可供计算机理解和处理的特征表示

- ## 时域特征

    直接从语音信号的时间波形中提取，计算简单

    - **短时能量**：反映语音信号的能量分布，用于区分有声段和无声段。
    - **过零率（ZCR）**：表示信号零点交叉的频率，适合区分清音和浊音
    - **音调周期**：用于分析语音的基频特性

- ## 频域特征

通过傅里叶变换分析信号的频谱特性

- **功率谱**：语音信号在频域上的能量分布
- **倒谱特征**：通过对对数功率谱进行傅里叶变换得到，以表示语音特征
- **共振峰**：由声道形状决定

- ## 时频域特征

  结合时间和频率特征，适用于非平稳信号的分析

  - **短时傅里叶变换（STFT）**：将信号分段后在每段上进行傅里叶变换，获得时频图
  - **梅尔频谱**：通过梅尔滤波器将频率域信息转换到梅尔频率域
  - **小波变换（CWT）**：对信号进行多尺度分析提取局部时频特征

- ## 声学特征

  基于人类听觉机制，即上述提到过的MFCC、PLP等

端到端的语音合成(主要是基于深度学习)

[@muReviewEndtoendSpeech2021]
[@LiHaiXiaJiYuDuanDaoDuanDeHanYuYuYinHeChengYanJiu2020]

- ## 语音合成发展历程（端到端的语音合成存在的必要性）

  传统语音合成经历了两个发展历程：单元挑选与波形拼接合成、参数合成

  - **单元挑选与波形拼接合成**：

    - 从标准音库中选取基元，经过边界调整后通过拼接方法得到完整的语音。
    - 优点：语音库为自然语音，合成的语音自然度较高
    - 缺点：对语音库的质量要求较高，生成的语音不具有鲁棒性，可扩展性差

  - **参数合成**

    - 用统计建模的方法，提取声学特征参数，并根据训练好的模型来预测声学特征。
    - **关键技术**
      - 隐马尔可夫模型（HMM）：一种概率模型，用来对非平稳信号进行建模
      - 高斯模型（GMM）：对声学特征参数进行建模
      - 深度神经网络（DNN）：一种深度前馈网络，能对文本和声学特征之间进行复杂建模

- ## 定义

  - 区别于传统的语音合成分为文本分析、特征提取、声码器等多个模块，端到端的语音合成通过一个统一的神经网络模型，从输入文本直接生成语音信号，跳过了多阶

段处理过程，从而使语音合成变得更加简明和快捷

- ## 工作原理

  - **文本处理**：文本首先被处理为可以输入模型的格式（梅尔频谱等）
  - **序列到序列的学习**：模型通过深度学习（如循环神经网络RNN，卷积神经网络CNN等），学习文本与语音信号之间的映射关系，将标准输入转换为中间声学特征（如频谱图、语音合成器特征等）
  - **语音生成**：将处理好的文本通过一个声码器（Griffin-Lim、WaveNet、HiFi-GAN等）转换为可以听到的音频

- ## 主要模型与技术

  - **Tacotron** [@wangTacotronEndtoEndSpeech2017]
    - 最早的提出端到端语音合成的模型之一，直接从字符合成语音。
  - **Tacotron2** [@eliasParallelTacotron22021]
    - Tacotron的扩展，可以高效地合成自然的语音。
  - **Fastspeech** [@renFastSpeechFastRobust]
    - 与Tacotron2等模型自回归生成梅尔频谱不同，Fastspeech提出了一种基于Transformer的新型前馈网络（FFT），实现了并行生成梅尔频谱图，并大大提高了生成速度；同时它运用了长度调节器，解决了音素不对齐导致的单词重复或跳过的问题，以及能够轻松调节语速。
  - **Fastspeech2** [@renFastSpeech2Fast2022]
    - Fastspeech2通过使用真实目标而非教师模型的简化输出来训练模型，从而解决了Fastspeech中流程复杂耗时的问题；同时引入了更多的语音变化信息作为条件输入，提高生成速度的同时也提升了语音质量。

- ## 作用与优势

  - **简化合成流程**：端到端模型通过将语音合成过程简化为一个神经网络模型，减少了传统方法中各个模块之间的复杂交互。
  - **提高语音质量**：端到端模型通过深度学习方法，能够捕捉到语音的细微差别，提高了生成语音的自然度与流畅性。
  - **加速开发与部署**：端到端模型通常具有较好的通用性，能够快速适应不同的应用场景，简化了模型的部署与维护。

- ## 缺点

  - **数据需求量大**：需要大量的高质量数据进行训练，数据中的噪声、标点符号等都会影响训练效果
  - **计算资源需求高**：端到端语音合成通常是基于深度学习的，其计算和学习都需要大量资源和成本

- ## 发展前景

- **多模态语音合成**：结合视觉、情感等多模态信息，使得语音合成更加个性化和情感丰富。例如，结合面部表情或情绪信息生成符合语境的语音。
- **低资源下的语音合成**：开发能够在低资源环境（如低计算能力、低训练数据）下运行的端到端语音合成模型，满足边缘计算、嵌入式设备等场景的需求。
- **实时与情感化合成**：随着计算能力的提升，端到端模型将能够以极低的延迟生成高质量的语音。同时，情感语音合成将为客服、虚拟助手等应用提供更加自然和有情感的互动体验。

神经声码器

[@AiYangMianXiangYuYinHeChengDeShenJingWangLuoShengMaQiYanJiu2021]
[@ZhangXiaoFeng]

- ## 定义

  - **声码器**：声码器是一种波形生成器，可以将处理好的声学特征转化为可供人耳听到的语音波形,是语音合成中重要的一部分，很大程度上关系着生成的语音是否流畅。
  - **神经声码器**：神经声码器可以基于深度学习，利用神经网络实现声码器的功能，使得生成的语音更加自然流畅。

- ## 分类

  - **自回归式**：

    - 按照时间顺序生成语音，每一部分生成的语音都依赖之前生成的语音，合成的质量较高。
    - **代表模型**：WaveNet、WaveRNN、SampleRNN

  - **并行式**：

    - **基于概率密度蒸馏**：
      - 用一个预先训练好的自回归式的教师模型训练一个基于**逆自回归流**的学生模型，生成的语音主要来自学生模型，由于学生模型没有自回归结构，因此生成效率更高
      - **代表模型**：Parallel WaveNet
    - **基于流**
      - 工作原理与基于概率密度蒸馏的声码器类似，但只需要一个模型便可完成，生成速度快。
      - **代表模型**：WaveGlow

  - **混合式（无自回归无流）**：

    - 结合自回归与并行的优势，不采用自回归使生成速度较快，不基于流使模型规模减小，并结合生成对抗网络（GAN）提高语音合成质量。
    - **代表模型**：HiFi-GAN、Parallel WaveGAN

- **主要模型与技术**
  - **WaveNet** [@oordWaveNetGenerativeModel2016]
    - “完全概率性且自回归”的深度神经网络，采用**自回归卷积神经网络**逐点生成音频波形，基于已生成的样本预测下一个点
    - **优点**：生成语音自然度较高，在MOS测试中取得了很高分数
    - **缺点**：计算成本较高，生成速度慢
  - **Waveglow** [@prengerWaveglowFlowbasedGenerative2019]
    - 基于流（glow），无需自回归，也无需使用教师-学生模型来训练。
    - **优缺点**：生成速度快，但生成质量较低
  - **HiFi-GAN** [@kongHiFiGANGenerativeAdversarial]
    - 以全卷积神经网络为生成器，结合CNN和GAN，生成速度很快，同时质量与自回归模型相当
- **发展趋势**：
  - **更高的合成质量**
  - **更快的推理速度**
  - **更低的计算成本**
  - **更强的鲁棒性**