**（1）**

大家好，我们组做的是端到端的语音合成，实现了两个功能：其一是根据输入的文本，在训练好的模型上生成对应的语音；其二是实现了实时语音克隆，可以在短时间内(几秒内)实现 输入目标说话人音频样本的音色克隆，并根据输入文本内容生成任意语音。

**（目录2）**

这两个功能的模型框架有共通之处，下面我将逐个展开介绍。

**（3）**首先介绍系统的原理

**（4）**功能一，即传统的端到端语音合成，是基于Tacotron 2的神经网络架构实现的。该系统由一个循环的 序列到序列 特征预测网络组成，该网络将文本序列映射到梅尔频谱；然后是一个经过修改的 WaveNet 模型作为声码器，将生成的梅尔频谱转化为时域上的波形，得到输出的音频。**（切5）**与tacotron类似，声学模型采用“编码器-注意力机制-解码器”结构，编码器将输入的文本序列转换成中间的特征序列；然后解码器结合基于位置敏感的注意力机制，生成预测梅尔频谱，输入给声码器。

**（6）**功能二，要实现实时语音克隆，方法是对训练有素的 TTS 模型进行条件化，从几秒钟语音音频中生成固定维度的嵌入向量，即提取出代表说话人音色的特征来，加入到TTS模型。参考SV2TTS结构，在实验一（Tacotron2）的基本结构上，增加说话人编码speaker encoder模块，由输入的参考语音生成嵌入，送入合成器与文本编码器的输出做拼接，再进行波形预测。这种方法通常比为每个说话者训练单独的 TTS 模型的数据效率更高，而且速度更快且计算成本更低

**（7）接下来，我们分别展示框架结构：**

**（8论文图）**这是参考论文提出的模型，（和刚才说到的模型是一样的）我从下到上简单地展开介绍一下系统的框架结构：

按照次序，先介绍encoder：

**（9 encoder）**

**结构：embedding + 3Conv\_layers + Bidirectional\_LSTM**

输入的文本，经过预处理成为字符序列，字符嵌入将每个character映射到512维向量，然后通过三层卷积层。这是由于实践中RNN很难捕获长时依赖，卷积层对长期上下文进行建模，并且能使得模型对不发音字符更为鲁棒(如'know'中的'k'。经过卷积，将被送入双向lstm，从而能感知到前后向的信息，实现上下文联系。

**（10 attention）**

而编码器生成的中间隐序列将被送入注意力网络：

location sensitive attention

这里使用位置敏感注意力机制，输入包括编码器隐状态、解码器隐状态、以及先前的注意力权重和累计注意力权重作为附加特征。这能鼓励模型随着预测的增加始终向前推进，实现注意力变化，从而有助于减少或避免解码器重复或忽略某些子序列的问题。H\_encoder经过新的注意力权重加权将被送入解码器**（11 换图片-解码器）**

**Decoder**

解码器一次一帧地从编码器的输入序列预测梅尔频谱。上一步预测首先通Pre-Net。论文指出：作为“信息瓶颈”的Pre-Net对于学习注意力至关重要。接下来，会通过 2个单向 LSTM 层，期间通过注意力网络读取编码器输入。再通过线性投影层，预测目标频谱图帧。最后经再经过post-net的残差叠加得到梅尔频谱。

同时，LSTM 的输出还会通过线性投影和sigmoid 激活来预测输出序列完成的概率，确定何时终止生成。

**12 WaveNet**

声码器负责将将梅尔频谱转化为波形，是一个相对独立的模块。论文采用改进的WaveNet。它采用膨胀卷积、因果卷积以扩大感受野，虽然参数量大训练时间相对长但质量较高，取得很好的效果。

**（13 SV2TTS-mocking bird）**

介绍了Tacotron2，我们来看实时语音克隆的实现框架：

可以看到合成器、声码器都是在原基础上实现的。重点要讲的就是speaker encoder模块。

**（14 只介绍speaker encoder）**

由于实现了说话人建模和语音合成解耦，speaker encoder可以单独进行说活人验证训练，采用GE2E损失函数。GE2E即广义端到端损失，它的训练是这样的·**（15）** x\_ij提取：将音频信号 以特定的宽和步长 分帧，然后我们提取 固定维数(大概40维)的对数梅尔滤波器组能量 作为每个帧的特征。w和b是可学习参数

**（16）**如图中的相似矩阵，我们希望彩色区域的相似值较大，而灰色区域的相似值较小

有两种实现方式**（17）**

Softmax损失函数意味 将每个嵌入向量推到其中心附近，并将其从所有其他中心拉出

(前一项-S\_ji,j拉近了positive pair之间的距离，后一项则对所有的negative pair一视同仁拉大它们之间的距离)

Contrast只增大那个相似度最大(即距离最近)的negative pair之间的距离。

得到每一段的embedding之后求平均得到最后的embedding

**part 3结果及分析：18**

训练模型：

**（19）**我们借助北大人工智能集群系统训练模型，使用标贝中文标准女声数据集、ljspeech数据集对编码器-解码器组合 以及 声码器分别展开训练；

这是在训练Tacotron2编码器和解码器的过程中的loss图像，**（20展示loss图）**随着训练步数的加深，引导对齐alignment逐渐形成清晰的对角线形态，表明模型逐渐开始收敛。在训练步数加深的过程中，我们取出模型进行测试，发现当训练到约200000步时，loss已经达到了基本满意的状态，推测接近极值/至少是局部极值，于是终止了训练；

右侧是第200000步最终得到的梅尔谱，最上面是目标频谱图，下面是我们生成的频谱，可以看出，已经比较接近了。

**（21放图）**这是论文中展示的可视化说话人编码结果，可以看到：左边，随着步数加深，来自相同说话人的样本被清晰地划分在一起；右边，10位说话人的嵌入有了清晰的区分……

**（22放图）**而这是我们为实现实时语音克隆过程中截取的两张图像，模型为上下两段样本生成了梅尔频谱图，并得出如左图所示的嵌入向量。

**（23听感反馈与改进措施）**

目前基本实现了对输入的短时间音频快速克隆的功能。

我们组内对系统进行测试，试听了生成语音，发现主要有两个地方需要改进：首先是，生成的语音混有噪声，在高频段有类似人声的伪影。未来可以通过对模型进行正则化，减少过拟合；同时，训练集中在解码器-编码器部分，未来可以进一步改进声码器，提高音频质量。

第二，做实时语音克隆的时候，如果说话人声音不大，或者有背景噪音，会影响到模仿的效果，使得生成的语音沙哑有噪声。针对这个问题，我们一个是对输入的音频进行预处理，进行降噪、语音增强，目前，把组员自己录制的音频经过软件降噪再输入，可以取得不错的效果（但是担心因此丢失说话人音色特征，所以把语音增强连接到编码器后对整体进行训练，将会取得更好的效果）。另一方面，可以加大说话人编码器在含噪样本上的训练量，提高特征提取能力。

**part 4**

**展示demo**