文章编号：1003-0530(2022)03-0527-09

一种基于改进注意力机制的实时鲁棒语音合成方法

唐 君 张连海 李嘉欣
（中国人民解放军战略支援部队信息工程大学信息系统工程学院，河南郑州 450001）

摘 要： 针对现有的语音合成系统Tacotron 2中存在的注意力模型学习慢、合成语音不够鲁棒以及合成语音速度较慢等问题，提出了三点改进措施：1. 采用音素嵌入作为输入，以减少一些错误发音问题；2. 引入一种注意力损失来指导注意力模型的学习，以实现其快速、准确的学习能力；3. 采用WaveGlow模型作为声码器，以加快语音生成的速度。在LJSpeech数据集上的实验表明，改进后的网络提高了注意力学习的速度和精度，合成语音的错误率相比基线降低了33.4%；同时，整个网络合成语音的速度相比之下提升约523倍，实时因子（Real Time Factor，RTF）为0.96，满足实时性的要求；此外，在语音质量方面，合成语音的平均主观意见分（Mean Opinion Score，MOS）达到3.88。

关键词： 语音合成；注意力损失机制；Tacotron 2；WaveGlow；序列到序列
中图分类号： TN912.33 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2022.03.010

引用格式： 唐君，张连海，李嘉欣. 一种基于改进注意力机制的实时鲁棒语音合成方法[J]. 信号处理，2022，38(3)：527-535. DOI：10.16798/j.issn.1003-0530.2022.03.010.

Reference format： TANG Jun，ZHANG Lianhai，LI Jiaxin. A real-time robust speech synthesis method based on improved attention mechanism[J]. Journal of Signal Processing，2022，38(3)：527-535. DOI：10.16798/j.issn.1003-0530.2022.03.010.

A Real-time Robust Speech Synthesis Method Based on Improved Attention Mechanism

TANG Jun ZHANG Lianhai LI Jiaxin
（School of Information System Engineering，PLA Strategic Support Force Information Engineering University，Zhengzhou，Henan 450001，China）

Abstract: In order to solve the problems of the existing speech synthesis system Tacotron 2，such as that the attention model is slow to learn，the synthesized speech is not robust enough，and the synthesized speech speed is slow，three improvement measures are proposed：1. Use phoneme embedding as input to reduce some mispronunciation problem；2. Introduce an attention loss to guide the learning of the attention model to realize its fast and accurate learning ability；3. Use the WaveGlow model as a vocoder to accelerate the speed of speech generation. Experiments on the LJSpeech data set show that the improved network improves the speed and accuracy of attention learning，and the error rate of its synthesized speech is reduced by 33.4% compared to the baseline；at the same time，the speed of synthesized speech of the entire network is increased by approximately 523 times，the Real-Time Factor（RTF）is 0.96，which meets the real-time requirements；in addition，in terms of voice quality，the Mean Opinion Score（MOS）of synthesized speech reaches 3.88.

Key words: speech synthesis；attention loss mechanism；Tacotron 2；WaveGlow；sequence to sequence

收稿日期：2021-06-11；修回日期：2021-07-30
基金项目：国家自然科学基金资助项目(61673395)

1 引言

语音合成[1](Speech Synthesis),也称文语转换(Text to Speech,TTS),是实现人机智能语音交互的关键技术之一。如今,随着语音合成技术的不断发展,其合成语音的质量不断得到提高,甚至已经达到与人类语音无法区分的程度。因此,语音合成的价值也不断受到人们的重视,其应用场景也十分广泛,例如智能家居、智能车载、智能客服、智能金融、智能教育、智能医疗等。

传统的TTS系统通常由前端和后端两个组件构成,前端负责文本分析和语言特征的提取,如包括分词、词性标注、多词消歧和韵律结构预测等功能;后端则基于前端的语言特征来合成语音,如包括语音声学参数建模、韵律建模以及语音生成等功能。在过去的几十年里,拼接语音合成[2-3]和参数语音合成[4-7]是主流技术,但这两种技术的处理流程都比较复杂,且需要事先人为定义好语言特征,而特定语言的语言特征的定义则需要更多的专家资源和人力资源。此外,这些方法合成的语音在韵律和发音方面往往存在“毛刺”、噪音或不稳定等现象,因此听起来不自然。

近年来,序列到序列(Sequence-To-Sequence,Seq2Seq)[8]的建模方法受到广泛关注,这类方法不仅在学习数据的固有特征方面有强大优势,而且还简化了传统语音合成方法的处理流程,仅利用单个模型就可以得到文本到声学特征之间的映射关系,这类方法也称为端到端语音合成。它们有许多优点:一是不需要定义复杂的语言特征,仅仅利用“<文本,音频>对”语料库进行训练,减轻了对特征工程的需求;二是单个模型的学习消除了流水线处理过程中存在的不兼容和误差累积等问题;三是利用神经网络可以学习更多的潜在属性,使合成的语音更自然,更具表达力。

如今,基于Seq2Seq的语音合成方法是主流技术,其合成语音的质量已经接近人类语音的质量,特别是基于注意力机制[9]的Seq2Seq模型在这一领域占有主导地位,如Tacotron 2[10]和Transformer TTS[11]等自回归模型。虽然这类模型合成语音的质量较高,但其鲁棒性不强,具体表现为对于复杂或特别难的句子,其合成的语音经常会存在跳词、重词、错词等现象,在这类模型中,Tacotron 2的鲁棒性要略好些,这得益于其采用位置敏感注意力机制[12]减少了一些错误模式。此外,也有Seq2Seq模型通过抛弃了注意力机制消除了合成的语音存在的跳词、重词、错词等现象,如FastPitch[13]、FastSpeech 2[14]、TalkNet 2[15]等非自回归模型,但这类模型由于没有注意力机制以学习文本到声学特征的对齐,因此,它们通常需要从一个事先训练好的外部模型(如Tacotron 2、Transformer TTS、语音识别模型等)中提取字素或音素的时长信息以训练一个时长预测网络来实现文本与声学特征之间的硬对齐。然而,这类模型想要达到与自回归模型相当的语音质量,通常需向文本中额外添加其他信息(如基频、能量等),这类模型训练过程复杂,工作量大。到目前为止,Tacotron 2仍然是语音合成领域中比较热门的模型,发展前景很大并且具有很好的通用性,利用它可以搭建很多其他任务的系统,比如语音克隆、语音风格控制、语音韵律控制、语码转换等。

Tacotron 2虽然具有不少优势,但是其仍然存在一些问题。首先,由于其采用字符嵌入作为模型输入,这会带来一些发音规则学习不到的问题,因为在不同的单词中相同字母的发音可能是不一样的,当训练语料库不够大时,不足以覆盖这些相同字母不同发音规则时,神经网络就学习不到这种规则,这会导致合成的语音有时会存在一些错误的发音。其次,它是个条件自回归模型,预测当前梅尔(Mel)谱需要前一时刻的Mel谱和上下文信息作为输入,然而在实际训练过程中,采用教师强迫的方式(即利用真实的Mel谱作为输入),导致训练和推理之间的不匹配,这种现象通常称为暴露偏差[16],它强化了条件自回归模型对局部信息的偏好,这会造成模型训练前期完全依靠教师强迫的输入而不使用上下文信息来预测Mel谱,又由于上下文信息由注意力机制计算产生,因此会导致注意力模块学习过程缓慢,并且注意力模块在学习的过程中可能会存在一些错误的文本到声学特征的对齐,这些错误对齐会导致合成语音出现跳词、重词、甚至胡言乱语。此外,Tacotron 2采用WaveNet[17]作为声码器,以将预测的Mel谱转化为语音波形,WaveNet的主体结构是一系列的膨胀卷积堆叠在一起,整个网络的感受野很大,由于其自回归结构导致其预测当前样本点,需要先前生成的样本点为条件,因此,逐样本点的生成机制造成其合成语音的速度十分缓慢,根本无法满足实时性的要求。由此可见,Tacotron 2存在三个明显的缺点:一是对于复杂或比较难的句子,其合成的语音容易出现错词、漏词、重词等错误;二是注意力模块的训练需要大量时间,并且其中的

“对齐学习”不够精确；三是合成语音速度很慢，无法实时合成。

针对上述存在的问题，本文提出了一些改进措施：1.为解决一些发音规则学习不到问题，采用音素序列替代字符序列作为输入；2.引入一种注意力损失来指导注意力模块的快速、准确的学习；3.采用并行的音频生成模型 WaveGlow[18]作为声码器，以替代原有的WaveNet声码器，以提高合成速度。

2　序列到序列模型

Seq2Seq模型是一类特殊的循环神经网络体系结构，其功能是实现一种序列$(x_1, x_2, \cdots, x_T)$到另一种序列$(y_1, y_2, \cdots, y_{T'})$转换。通常来说，输入序列和输出序列的长度一般不相等，即$T \neq T'$，并且每个y_t的预测都是以先前所有预测输出$(y_1, y_2, \cdots, y_{t-1})$为条件，因此输入序列到输出序列的转换关系可用条件概率$p(y_1, y_2, \cdots, y_{T'}|x_1, x_2, \cdots, x_T)$进行表示。

Seq2Seq模型最常见的架构是编码器-解码器(encoder-decoder)结构，该架构包含两个组件：编码器(encoder)和解码器(decoder)。目前大多数的编码器-解码器结构都结合了注意力机制，注意力机制的引入极大提升了编码器-解码器的性能。编码器和解码器的作用可分别由式(1)、(2)表示：

$$h_t = \text{encoder}(h_{t-1}, x_t) \tag{1}$$

$$s_t = \text{decoder}(s_{t-1}, y_{t-1}, c_t) \tag{2}$$

其中，h_t、h_{t-1}分别表示编码器的当前时间步的隐藏状态和上个时间步的隐藏状态；s_t、s_{t-1}分别表示解码器的当前时间步的隐藏状态和上个时间步的隐藏状态。c_t表示当前时间步的上下文向量，其由注意力模块计算，如式(3)表示：

$$c_t = \text{attention}(s_{t-1}, \boldsymbol{h}) \tag{3}$$

条件概率$p(y_1, y_2, \cdots, y_{T'}|x_1, x_2, \cdots, x_T)$可进行分解，如式(4)所示：

$$p(y_1, y_2, \cdots, y_{T'}|x_1, x_2, \cdots, x_T) = \prod_{t=1}^{T'} p(y_t|y_{<t}, \boldsymbol{x}) \tag{4}$$

而式(4)中的概率项$p(y_t|y_{<t}, \boldsymbol{x})$可利用式(2)中的隐含状态$s_t$进行建模，如式(5)所示：

$$p(y_t|y_{<t}, \boldsymbol{x}) = \text{softmax}(f(s_t)) \tag{5}$$

其中，$f(\cdot)$表示一个全连接层。

对于机器翻译[19]任务，常用softmax函数来计算词汇表中每个单词的概率。然而，在TTS任务中，解码器计算的隐藏状态$\boldsymbol{s}$直接通过一层线性映射层预测得到目标Mel谱，因此并不需要softmax函数。

3　基于Tacotron 2的改进模型

Tacotron 2是一个典型的结合注意力机制的序列到序列模型，也是目前热门的语音合成模型之一，但它仍然存在一些关键性问题。本文针对这些问题，对Tacotron 2模型进行了改进，改进后的模型框架如图1所示。

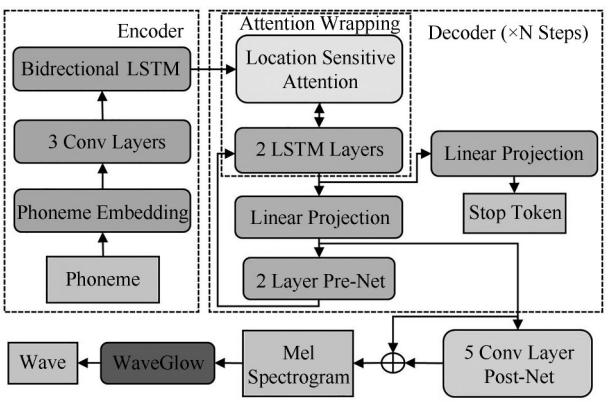


图1　基于Tacotron 2的改进结构原理图

Fig. 1　Schematic diagram of the improved structure based on Tacotron 2

整个模型由特征预测网络和声码器网络两个部分构成。本文在Tacotron 2原有的结构基础上将字符嵌入换为音素嵌入，同时在特征预测网络中引入注意力损失来指导注意力模块的学习；在声码器网络部分，将原有的WaveNet网络替换为WaveGlow网络，模型的具体细节将在以下小节中叙述。

3.1　文本到音素的转换

英语语音存在相同字母发音不同的现象，这些复杂的发音规则，需要大量的训练数据才能学习到，因此在利用神经网络学习这些发音规则时，若训练数据不够充足，就很难学习到所有的发音规则。尤其当某些规则在训练数据中出现次数太少时，神经网络将无法充分学习到这些规则，必然会导致某些合成的语音出现发音错误问题。而音素(类似于中文中的声母、韵母)作为发音的最基本的单元，本身就体现了发音属性，因此以音素序列作为输入，可避免发音错误问题。音素序列利用开源的字素到音素转换工具G2P从文本中提取，标点符号也作为一种特殊的音素序列包括在内，因为标点符号也携带着一些信息，如逗号通常在语音中体现为停顿。

3.2 Encoder

编码器负责将音素序列转换为隐藏特征表示。本文使用可学习的512维音素嵌入来表示输入音素序列，编码器首先利用3层的卷积层对输入音素序列的长期上下文信息进行建模，每层卷积层包括512个形状为5 × 1的卷积核（每个卷积核跨越5个音素），并且每层卷积层跟随有修正线性单元（Rectified Linear Unit，ReLU）激活和批归一化（Batch Normalization，BN）处理。最后一个卷积层的输出被送入到一层包含512个单元（每个方向256个）的双向长短时记忆（Bi-directional Long Short-Term Memory，BiLSTM）网络中，以生成隐藏特征表示，即一段话中每个音素序列都被重新编码为512维的特征表示。

3.3 Decoder

解码器是一个自回归循环神经网络，负责利用编码器生成的隐藏特征来进行Mel谱预测。编码器的输出首先被送入注意力网络（对应于图1中Attention Wrapping的部分）中，注意力网络具体结构如图2所示，该注意力网络负责将编码器的输出转化为固定长度的上下文向量。这里注意力网络采用位置敏感注意力机制，它扩展了附加注意力机制[8]，使用来自先前解码器时间步长的累积注意力权重作为附加特征，这促使注意力模型在输入过程中一致地向前移动，从而减轻了一些子序列被解码器重复或忽略的潜在错误模式。

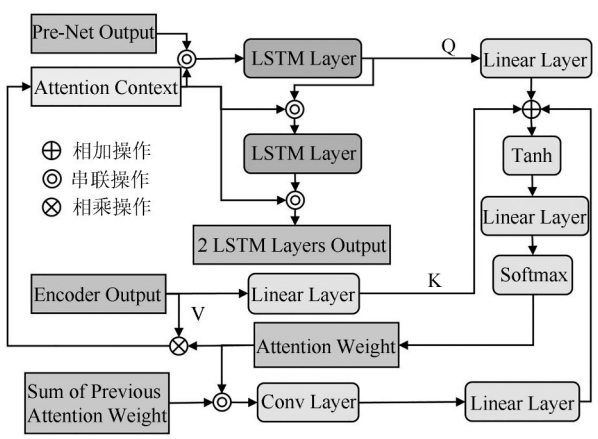


图2 注意力网络结构图

Fig. 2 Attention network structure diagram

在训练时，前一个时间步的真实Mel谱（推理阶段采用预测的Mel谱）被输入到具有ReLU激活的两层全连接层中，称其为解码器预网，它作为一个信息瓶颈层，在整个系统中起着很重要的作用。因为音素具有可训练的嵌入性，因此它们的子空间是自适应的，而Mel谱的子空间是固定的，解码器预网负责将Mel谱映射到与音素嵌入相同的子空间中，这样才能更有效的计算<音素，Mel谱>之间的相似性，从而发挥注意力机制的作用。

如图2所示，前一个时间步的Mel谱帧经过预网的输出与前一个时间步的上下文向量串联在一起送入第一层LSTM（1024个单元）中，其输出称为查询（Query，Q）向量。Q向量经过一层线性层（128个单元）处理，其输出称为处理后的Q向量；与此同时，Q向量与前一个时间步的上下文向量串联输入到第二层LSTM（1024个单元）中生成解码器隐状态，解码器隐状态再与前一个时间步的上下文向量串联在一起作为注意力模块当前时间步的输出，即图1中2层LSTM的输出。这里将编码器输出称为值（Value，V）矩阵（矩阵中第i个行向量对应于一句话中第i个音素序列经过编码器编码的隐含特征表示），V矩阵通过一层线性层（128个单元）生成键（Key，K）矩阵。前一个时间步的注意力权重与先前所有的注意力权重的累积串联在一起送入一层卷积层中，通过32个长度为31的一维卷积核计算生成位置特征。位置特征通过一层线性层（128个单元）处理后作为附加特征与K矩阵和处理后的Q向量一起经过Tanh函数处理，再经过一层线性层以生成输出，称其为注意力概率向量（该概率向量中的第j个分量的大小代表Q向量与K矩阵第j个行向量的相关度的大小）。最后通过softmax函数处理注意力概率向量得到当前注意力权重向量，当前注意力权重向量与V矩阵相乘得到当前上下文向量。

注意力模块当前时间步的输出分别送入两个不同的线性层中，一个线性层利用其输出预测当前Mel谱帧，另一个线性层则将其转化为一个标量，并传递给sigmoid函数来预测输出序列完成的概率。在推理过程中，使用这个“停止令牌”预测（输出概率大于0.5时，继续下一步解码过程；反之，停止解码过程），以允许模型动态地决定何时终止预测Mel谱，而不是在固定时间内终止预测。最后，通过N次解码步骤（第N次达到停止解码的条件）得到预测的Mel谱将被送入后处理网络中，产生一个残差与自身叠加生成最终的Mel谱，以改善整体重建。后处理网络由5层的卷积层构成，每层卷积由512个形状为5 × 1的卷积核组成，并且每层卷积层后接BN层，除最后一层卷积层外，其他四层都采用Tanh函数激活。

3.4　损失函数

在原始的Tacotron 2中，其损失函数由两部分组成：一是解码输出的Mel谱和经过后处理网络处理的Mel谱分别与目标Mel谱的最小均方误差损失；二是预测终止条件的二值交叉熵损失。由此可见，整个模型的注意力模块的学习过程是网络自行学习的，并没有进行干预。

在实际训练过程中，注意力模块的学习需要耗费大量的训练时间，并且可能学习到一些错误对齐行为，受DCTTS[20]中的引导注意力的思想启发，本文采取向Tacotron 2中的注意力模块引入损失，以作为先验知识来指导注意力模块的学习，保证其能更好的学习<音素，Mel谱>之间的对齐关系，准确的对齐将在每次步骤过程中提供更准确的上下文向量，这有助于间接缓解条件自回归模型因教师强迫训练而带来的局部信息偏好问题，从而减少合成的语音中出现的漏词、重词等错误行为，使得整个系统更鲁棒。

对于一段语音来说，其字符串或音素串与音频片段的顺序是基本一致的，换句话来说，当人们朗读一句话时，很自然假设文本的位置n与发音时间t成线性关系。基于这种规则，在语音合成中，理想的注意力矩阵应该是一种类似对角矩阵的结构。这也是语音合成技术与其他SeqSeq技术的显著区别，比如在机器翻译中，注意力模块只需要解决具有不同语法规则的两种语音之间的单词对齐。

本文通过引入一种注意力损失来引导注意力矩阵$\boldsymbol{A}$向着对角矩阵靠拢，注意力损失定义如式(6)所示。

$$\mathcal{L}_{\text{attention}} = \sum_{n=0}^{N-1}\sum_{t=0}^{T-1} A_{nt}W_{nt} \tag{6}$$

其中，A_{nt}是注意力矩阵$\boldsymbol{A}$中的一个元素，即代表第$n+1$个字符串或音素串与第$t+1$个时间片段的对齐概率值，N、T分别指一段语音中字符串或音素串的总个数，时间片段的总长度(Mel谱的总帧数)，W_{nt}是引导矩阵$\boldsymbol{W}$中的第$n+1$行、第$t+1$列的一个元素，定义如式(7)所示。

$$W_{nt} = 1 - \exp\left\{-\left(\frac{n}{N}-\frac{t}{T}\right)^2 \Big/ 2g^2\right\} \tag{7}$$

其中，g是权重因子，引导注意力矩阵向对角矩阵靠拢的程度。

一个引导矩阵的示例如图3所示，当注意力矩阵$\boldsymbol{A}$远离对角线，例如以随机顺序读取字符串或音素串，它将受到注意力损失函数的强烈惩罚，虽然

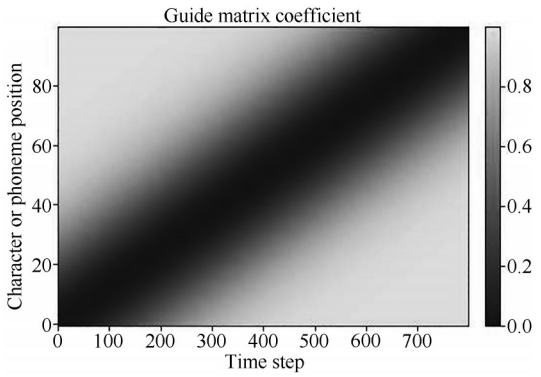


图3　引导矩阵($g=0.2, N=100, T=800$)

Fig. 3　Guide matrix($g=0.2, N=100, T=800$)

上述的假设是不够充分的，但是它将有助于注意力模型快速的学习对齐，并减少一些错误学习行为。

本文通过将上述的注意力损失引入Tacotron 2中来指导其注意力学习对齐过程，这里的注意力损失$\mathcal{L}_{\text{attention}}$与Tacotron 2原有的损失$\mathcal{L}_{\text{original}}$共同优化，整个特征预测模型的损失函数如式(8)所示。

$$\mathcal{L} = \mathcal{L}_{\text{original}} + \lambda\mathcal{L}_{\text{attention}} \tag{8}$$

其中，λ控制注意力损失的相对权重，实验中λ的值被设置为100。

3.5　WaveGlow声码器

本文采用WaveGlow作为声码器网络，以将Mel谱转换为语音波形。WaveGlow是一种基于流的模型，它借鉴了图像生成模型Glow[21]和WaveNet的优点，但不同于WaveNet的逐样本点自回归式的音频生成，WaveGlow可以并行生成高质量的音频样本。在训练期间，WaveGlow将输入语音波形x转换为零均值球面高斯分布z，相反，在推理期间，通过逆运算从零均值球面高斯分布z中随机采样生成语音波形x，过程如式(9)、式(10)、式(11)所表示。

$$z \sim N(z;\ 0, \boldsymbol{I}) \tag{9}$$

$$x = f_0 \circ f_1 \circ \cdots \circ f_k(z) \tag{10}$$

$$z = f_k^{-1} \circ f_{k-1}^{-1} \circ \cdots \circ f_0^{-1}(x) \tag{11}$$

其中，$\boldsymbol{I}$代表单位矩阵，f_i代表第i个变换，f_i^{-1}代表第i变换的逆变换，$f_i \circ f_j(\cdot)$代表$f_i(f_j(\cdot))$。

因此，WaveGlow由一系列变换组成，以逐步将语音数据映射到高斯空间，变换由可逆1×1卷积[21]和放射耦合层[22]组成，通过直接最小化数据的负对数似然度来训练模型。以Mel谱特征作为条件，模型最终的负对数似然度公式如式(12)所示。

$$-\log p_\theta(x)=\frac{z(x)^T z(x)}{2\sigma^2}-\sum_{j=0}\log s_j(x,h_{\text{mel}})-\sum_{k=0}\log\det\left|W_k\right| \tag{12}$$

其中，s_j、W_k、σ^2、h_{mel}分别代表仿射耦合层中第j个WaveNet网络的输出系数、可逆1×1卷积的第k个加权矩阵、零均值球面高斯分布的假设方差和Mel谱特征。

4 实验及结果

4.1 实验设置

实验数据采用公开的LJSpeech数据集，该数据集包括13100个英语音频片段和相应的文本，音频总长度约为24个小时，采样率为22050 Hz，由一名专业的女性录制。实验中将数据集随机分成两部分：12800个音频样本用于训练集，300个音频样本用于测试集。实验在单个GPU（NVIDIA GeForce GTX 1080Ti，内存为11 GB）上进行，实验中语音合成系统架构基于PyTorch搭建，系统中采用80维的Mel谱作为声学特征，其中FFT长度、帧长和帧移分别设置为1024、1024和256个采样点。

4.2 模型配置

特征预测网络和声码器网络是单独训练的，训练数据均采用LJSpeech数据集。训练特征预测网络，采用教师强迫模式训练，即每个预测帧以输入的音素序列和真实Mel谱的前一帧作为条件，实验中批处理大小设置为32，采用ADAM优化器，其中$\beta_1=0.9$、$\beta_2=0.999$、$\varepsilon=1e^{-6}$，学习率设置为$2e^{-3}$。

采用真实的Mel谱训练WaveGlow声码器，其模型参数与文献中相同。批处理长度和批处理大小分别为16000个样本点和8个样本，训练和推理时假设方差$\sigma=1$。为进行对比实验，本文同时训练了一个WaveNet网络，网络参数保持与文献中相同，批处理长度和批处理大小分别为16000个样本点和8个样本。

4.3 注意力学习评估

为评估注意力损失在Tacotron 2中的有效性，本文分别在不同的训练阶段测试模型的注意力学习情况，如图4所示。引入注意力损失的模型仅仅通过10个训练周期，就出现了明显的对齐效果；50个训练周期后，开始显现清晰的对齐效果；100个周期后，模型已经学习到不错的注意力对齐效果（注意力对齐更集中且对齐概率值更大），与输入的文本进行比对，可以看出模型已经捕获了一些节奏信息，如图4右下角子图中的两个红圈的位置对应于文本中的逗号，即对齐的是静音帧。仅利用经过100个周期训练后的模型来合成语音，其合成的语音已经十分清晰，逗号带来的停顿感也十分明显。

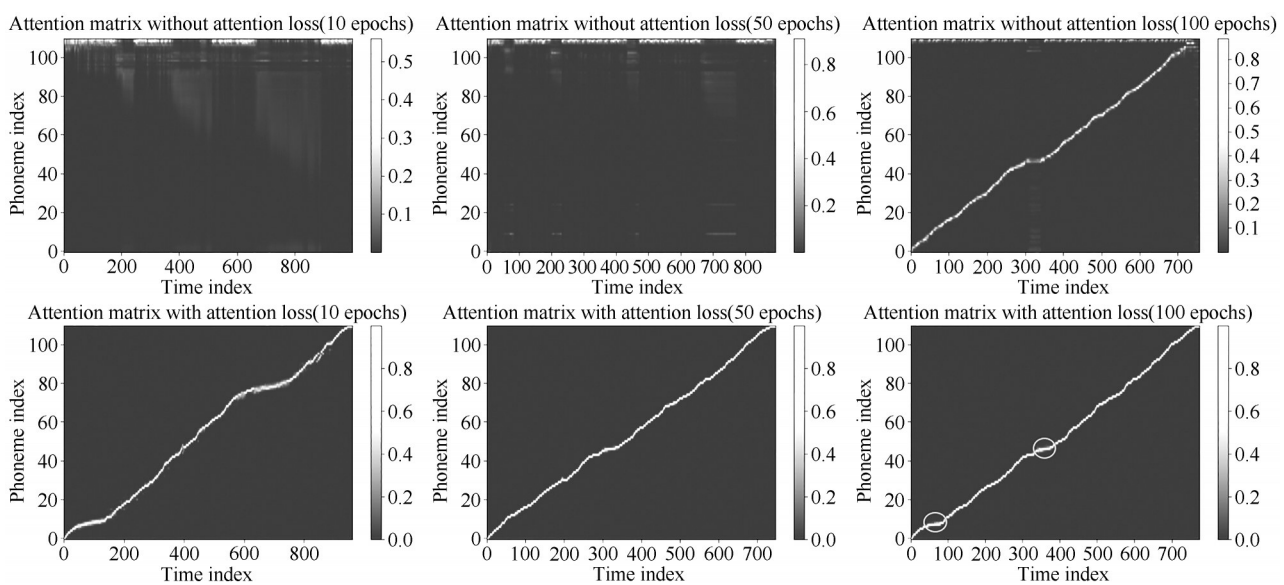


图4 不同阶段的注意力矩阵图（从测试集随机挑选的一段话，其文本内容为“Printing, in the only sense with which we are at present concerned, differs from most if not from all the arts and crafts represented in the Exhibition”）

Fig. 4 Attention matrix at different stages (a paragraph ran-domly selected from the test set, the text content is "Printing, in the only sense with which we are at present concerned, differs from most if not from all the arts and crafts represented in the Exhibition")

在训练过程中，发现通过引入注意力损失不仅能加快注意力模块的学习过程，甚至加快整个模型的训练过程，而且模型损失相比原来反而略微下降。如图 5 所示，未加入注意力损失的模型损失约在 350 Epochs 左右开始趋于稳定，但是加入注意力损失的模型在 500 Epochs 时仍呈现略微下降的趋势，表明模型损失仍存在下降的空间。出现这种现象的原因，一是注意力损失指导注意力模块产生更准确的对齐，从而生成更准确的上下文向量，因此解码预测的更精准；二是注意力损失加速了注意力模块的学习过程，因此，其损失下降速度比原始模型要略快。

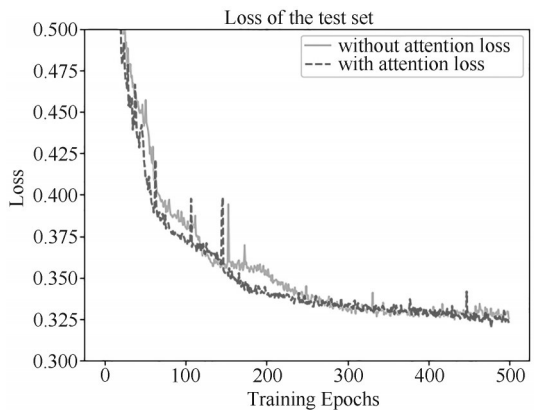


图 5　不同的训练阶段在测试集上的损失

Fig. 5　The loss on the test set in different training stages

可见，引入注意力损失确实能很好的指导注意力模型的学习，既能加快其学习过程，也能指导其学习正确的对齐。

4.4　系统鲁棒性评估

自回归模型中编码器-解码器的注意力机制可能会导致音素和 Mel 谱之间的错误对齐，从而导致单词出现重复和跳跃等不稳定现象。为评估音素嵌入和注意力损失的引入对 Tacotron 2 鲁棒性的影响，本文参考 FastSpeech[23] 中采用的 50 个对 TTS 系统来说特别难的句子（这些句子均不在 LJSpeech 数据集中），从中挑选 30 个句子来测试不同条件下的模型合成的语音内容的准确性，为了公平起见，模型均训练到拟合。

发音错误情况统计如表 1 所示，原 Tacotron 2 系统对这些句子的鲁棒性不强，错误率高达 36.7%，通过引入音素嵌入和注意力损失分别减少了 10% 和 23.4% 的错误率。当这两者结合，系统合成语音的错误率低至 3.3%，其鲁棒性与非自回归模型 FastSpeech 相接近。

表 1　不同条件下的系统在 30 个特别难的句子上的鲁棒性对比（同一句话中相同的错误只统计一次）

Tab. 1　Robustness comparison of the system under different conditions on 30 particularly difficult sentences (the same error in the same sentence is only counted once)

模型	重词数	漏词数	错词数	总错误句子	错误率
Transformer TTS	6	10	7	14	46.7%
Tacotron 2（原始）	5	8	4	11	36.7%
Tacotron 2 + 音素嵌入	4	8	1	8	26.7%
Tacotron 2 + 注意力损失	1	0	3	4	13.3%
FastSpeech	0	0	0	0	0.0%
本文的方法	0	0	1	1	3.3%

实验中，存在一个有趣的现象：前四个模型对单个字母均表现为无法合成正确的发音，并且均不能自行停止生成 Mel 谱，直至解码到最大解码步骤后停止生成。而本文的方法能够很好的解决了这种现象，如其合成字母 R 的注意力矩阵和 Mel 谱如图 6 所示，数据集中是没有单独字母 R 的发音的，可

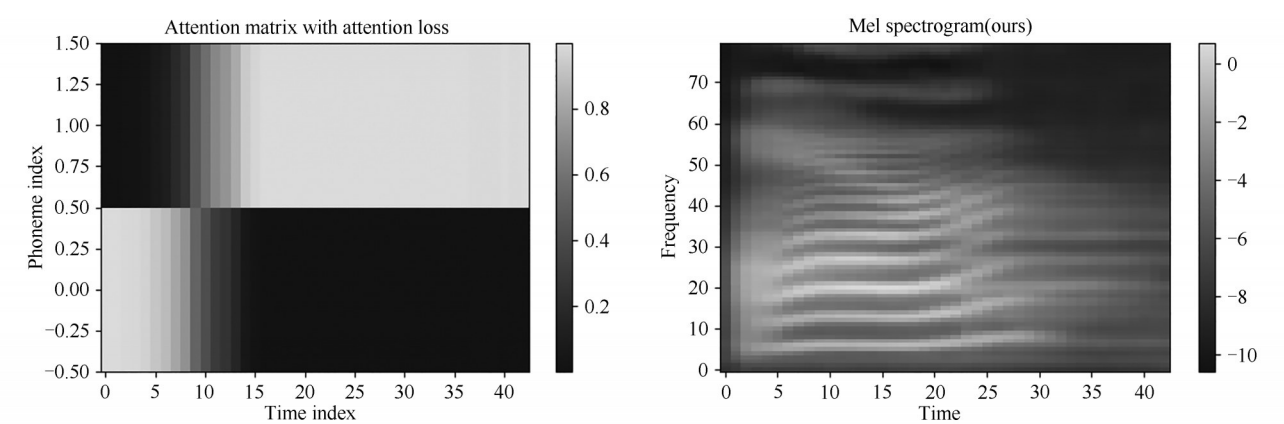


图 6　字母 R 的注意力矩阵和 Mel 谱图（字母 R 的音素为“AA1”和“R”）

Fig. 6　The attention matrix and Mel spectrum of the letter R (the phonemes of the letter R are "AA1" and "R")

以认为字母 R 作为单个单词是集外词，可以看出改进后的模型对集外词具有不错的适用性。

以上实验表明，采用音素作为输入能够减少 Tacotron 2 合成中存在的一些错误发音问题，注意力损失可以缓解其合成中存在的漏词和重词等现象。将这两者结合产生了更好的效果，极大提升了 Tacotron 2 系统的鲁棒性，提高了合成语音的准确度。这主要是因为音素的引入不仅消除了不同单词中同一个字母的发音不同问题，而且使得注意力模块的学习目标从原来的字素到 Mel 谱的对齐转变为音素到 Mel 谱的对齐，减少了注意力模块对齐学习的复杂度。同时注意力损失的引入提高了注意力模块的快速准确的学习能力，即提高了音素与 Mel 谱的对齐学习速度和对齐概率值，减轻 Tacotron 2 解码预测当前 Mel 谱时对上一帧的 Mel 谱过度依赖，使其侧重结合上下文信息来进行解码预测。

4.5 推理速度评估

原始的 Tacotron 2 模型采用 WaveNet 作为声码器，将预测的 Mel 谱转换成语音波形，但由于其预测当前样本是基于之前生成的样本点为条件，逐样本点生成的机制导致其合成语音的速度很慢，无法满足实时性的要求。针对这一问题，本文采用 WaveGlow 作为替换，由于 WaveGlow 能在推理阶段并行生成语音样本点，因此其生成语音的速度非常快。在实验中进行了本文的模型与原始的 Tacotron 2 的推理速度（从输入文本到输出语音所需的时间损耗）比较，采用实时因子（Real Time Factor，RTF）来衡量推理速度，RTF 表示系统合成一秒波形所需的时间（包括文本转换为音素的时间，单位秒），结果如表 2 所示。

表 2 合成语音速度比较

Tab. 2 Comparison of synthesized speech speed

系统	推理速度（RTF）
Mel 谱 + WaveNet	501.61
Mel 谱 + WaveGlow	0.21
Tacotron 2 + WaveNet	502.35
本文的方法	0.96

实验表明，通过引入 WaveGlow 作为声码器，极大提升了整个系统的合成速度，其 RTF<1，满足实时性的要求。这里可以看出 WaveNet 将 Mel 谱转化为语音波形的时间过慢是影响 Tacotron 2 合成语音速度的重要原因。

4.6 语音质量评估

本文采用平均主观意见分（Mean Opinion Score，MOS）来衡量合成语音的质量，从测试集中随机选取 30 条音频作为评估集，由 15 位精通英语的听众通过耳机试听给出主观评分，根据语音的质量，由差（1）到好（5）采用 5 分制进行打分，不同模型的 MOS 分值如表 3 所示，置信区间为 95%。

表 3 不同模型的 MOS

Tab. 3 MOS of different models

模型	MOS
真实语音	4.53 ± 0.07
真实 Mel 谱 + WaveNet	4.15 ± 0.08
真实 Mel 谱 + WaveGlow	4.17 ± 0.08
Tacotron 2 + WaveNet	3.69 ± 0.09
FastSpeech + WaveGlow	3.70 ± 0.08
Transformer TTS + WaveGlow	3.80 ± 0.09
Tacotron 2 + WaveGlow	3.78 ± 0.10
Tacotron 2（音素嵌入）+ WaveGlow	3.81 ± 0.07
Tacotron 2（注意力损失）+ WaveGlow	3.84 ± 0.08
本文的方法	3.88 ± 0.09

实验结果表明，采用音素嵌入和注意力损失提高了 Tacotron 2 模型合成语音的质量。最后，本文通过结合高效的音频生成模型 WaveGlow 作为模型的声码器，整个系统合成语音的 MOS 达到 3.88，相比原始模型的 MOS 提升了 0.19。

5 结论

本文分析了 Tacotron 2 中存在的问题，提出了三点改进措施：1. 采用音素嵌入作为网络输入；2. 引入注意力损失；3. 采用 WaveGlow 作为声码器。在 LJSpeech 数据集上的实验表明，改进后的 Tacotron 2 模型对复杂或难的句子存在的错误发音问题得到缓解，系统的鲁棒性得到极大提高；其次，注意力模块在学习<音素，Mel 谱>对齐的速度和准确度得到明显提升；最后，整个模型合成语音的速度满足实时性的要求，合成的语音更自然，语音质量也令人满意。

参考文献

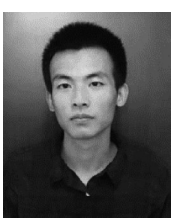
[1] SPROAT R W, OLIVE J P. Text-to-speech synthesis[J]. AT&T Technical Journal, 1995, 74(2): 35-44.

[2] HUNT A J, BLACK A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP). Atlanta, USA: IEEE, 1996: 373-376.

[3] BLACK A W, TAYLOR P. Automatically clustering simi-

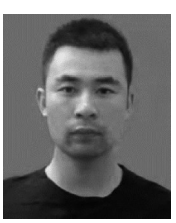
lar units for unit selection in speech synthesis[J]. Eurospeech, 1997, 2(5):601-604.
[4] ZEN H, TOKKUDA K, BLACK A W. Statistical parametric speech synthesis[J]. Speech Communication, 2009, 51(11):1039-1064.
[5] TOKDA K, YOSHIMURA T, MASUKO T, et al. Speech parameter generation algorithms for HMM-based speech synthesis [C]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Istanbul, Turkey:IEEE,2000:1315-1318.
[6] ZEN H, SENIOR A, SCHUSTER M. Statistical parametric speech synthesis using deep neural networks[C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). BC, Canada: IEEE, 2013:7962-7966.
[7] TOKUDA K, NANKAKU Y, TODA T, et al. Speech synthesis based on hidden markov models[J]. Proceedings of the IEEE, 2013, 101(5):1234-1252.
[8] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [EB/OL]. 2014: arXiv: 1409.3215[cs.CL]. https://arxiv.org/abs/1409.3215.
[9] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. 2014: arXiv: 1409.0473[cs.CL]. https://arxiv.org/abs/1409.0473.
[10] SHEN J, PANG R, WEISS R J, et al. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). AB, Canada: IEEE, 2018:4779-4783.
[11] LI N, LIU S, LIU Y, et al. Neural speech synthesis with transformer network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1):6706-6713.
[12] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[J]. Computer Science, 2015, 10(4):429-439.
[13] ŁAŃCUCKI A. Fastpitch: Parallel text-to-speech with pitch prediction[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ON, Canada:IEEE, 2021: 6588-6592.
[14] REN Y, HU C, TAN X, et al. Fastspeech 2: fast and high-quality end-to-end text to speech[EB/OL]. 2020: arXiv: 2006.04558[eess.AS]. https://arxiv.org/abs/2006.04558.
[15] BELIAEV S, GINSBURG B. TalkNet 2: non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction [EB/OL]. 2021: arXiv: 2104.08189[eess.AS]. https://arxiv.org/abs/2104.08189.
[16] RANZATO M A, CHOPRA S, AULI M, et al. Sequence level training with recurrent neural networks [EB/OL]. 2015: arXiv: 1511.06732[cs.LG]. https://arxiv.org/abs/1511.06732.
[17] OORD A, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[EB/OL]. 2016: arXiv: 1609.03499 [cs.SD]. https://arxiv.org/abs/1609.03499.
[18] PRENGER R, VALLE R, CATANZARO B. WaveGlow: a flow-based generative network for speech synthesis [C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019:3617-3621.
[19] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation [EB/OL]. 2016: arXiv: 1609.08144[cs.CL]. https://arxiv.org/abs/1609.08144.
[20] TACHIBANA H, UENOYAMA K, AIHARA S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). AB, Canada:IEEE, 2018: 4784-4788.
[21] KINGMA D P, DHARIWAL P. Glow: generative flow with invertible 1x1 convolutions[EB/OL]. 2018: arXiv: 1807.03039[stat.ML]. https://arxiv.org/abs/1807.03039.
[22] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real nvp[EB/OL]. 2016: arXiv: 1605.08803[cs.LG]. https://arxiv.org/abs/1605.08803.
[23] REN Y, RUAN Y, TAN X, et al. FastSpeech: fast, robust and controllable text to speech[C]// Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS, 2019:3171-3180.

作者简介



唐　君　男,1996年生,江西九江人。中国人民解放军战略支援部队信息工程大学硕士研究生,主要研究方向为智能信息处理、人工智能、语音合成。
E-mail: 2433548528@qq.com

张连海　男,1971年生,山东单县人。中国人民解放军战略支援部队信息工程大学教授,主要研究方向为语音信号处理、智能信息处理、人工智能、信号分析等。
E-mail: llhhzz163@163.com



李嘉欣　男,1998年生,湖南湘乡人。中国人民解放军战略支援部队信息工程大学硕士研究生,主要研究方向为智能信息处理、人工智能、语音合成。
E-mail: 414171817@qq.com