

文章编号: 1000-5641(2019)04-0111-09

## 一种基于 Tacotron 2 的端到端中文语音合成方案

王国梁<sup>1</sup>, 陈梦楠<sup>2</sup>, 陈 蕾<sup>2</sup>

(1. 国家电网安徽省电力有限公司 信息通信分公司, 合肥 230061;  
2. 华东师范大学 计算机科学技术系, 上海 200062)

**摘要:** 颠覆性设计的端到端语音合成系统 Tacotron 2, 目前仅能处理英文。致力于对 Tacotron 2 进行多方位改进, 设计了一种中文语音合成方案, 主要包括: 针对汉字不表音、变调和多音字等问题, 添加预处理模块, 将中文转化为注音字符; 针对现有中文训练语料不足的情况, 使用预训练解码器, 在较少语料上获得了较好音质; 针对中文语音合成急促停顿问题, 采用对交叉熵损失进行加权, 并用多层感知机代替线性变换对停止符进行预测的策略, 获得了有效改善; 另外通过添加多头注意力机制进一步提高了中文语音合成音质。梅尔频谱、梅尔倒谱距离等的实验对比结果表明了方案的有效性: 可以令 Tacotron 2 较好地适应中文语音合成的要求。

**关键词:** 语音合成; 多头注意力; Tacotron 2

**中图分类号:** TP391    **文献标志码:** A    **DOI:** 10.3969/j.issn.1000-5641.2019.04.011

## An end-to-end Chinese speech synthesis scheme based on Tacotron 2

WANG Guo-liang<sup>1</sup>, CHEN Meng-nan<sup>2</sup>, CHEN Lei<sup>2</sup>

(1. *Information and Communication Branch, State Grid Anhui Electric Power Co., Ltd., Hefei 230061, China;*  
2. *Department of Computer Science and Technology, East China Normal University, Shanghai 200062, China*)

**Abstract:** The disruptively design for an end-to-end speech synthesis system Tacotron 2, is currently only available in English. This paper is devoted to implementing several improvements to Tacotron 2 and presents a Chinese speech synthesis scheme, including: a pre-processing module to convert Chinese characters into phonetic characters to address the challenge of Chinese character not corresponding to pronunciation, having multiple tones, and having polyphonic words; a pre-training decoder to achieve better sound quality with less corpus given the lack of existing Chinese training corpus; a strategy of weighting the cross-entropy loss and using the multi-layer perceptron, instead of the linear transformation, to predict stop tokens and to solve the Chinese speech synthesis sudden pause problem; and a multi-head attention mechanism to further improve Chinese speech quality. The experimental comparison of the Mel spectrum and the Mel cepstrum

---

收稿日期: 2018-10-28

第一作者: 王国梁, 男, 硕士, 高级工程师, 长期从事电力信息化建设和电力信息化管理工作。

通信作者: 陈 蕾, 女, 博士, 副教授, 研究方向为计算机网络与智能系统。

E-mail: lchen@cs.ecnu.edu.cn.

distance (MCD) shows that our work is effective and can make Tacotron 2 adapted to the requirements of Chinese speech synthesis.

**Keywords:** text to speech; multi-head attention; Tacotron 2

## 0 简 介

语音合成,又称文语转换(Text To Speech, TTS),是一种可以将任意输入文本转换成相应语音的技术。传统语音合成系统通常包括前端和后端两个部分。前端主要对输入文本进行分析,提取某些语言学信息;中文合成系统的前端部分一般包含文本正则化、分词、词性预测、多音字消歧、韵律预测等模块<sup>[1]</sup>. 后端则通过一定方法,例如参数合成或拼接合成、生成语音波形。

参数合成指基于统计参数建模的语音合成<sup>[2]</sup>. 该方法在训练阶段对语言声学特征、时长信息进行上下文相关建模,在合成阶段通过时长模型和声学模型预测声学特征参数,对声学特征参数做后处理,最终利用声码器恢复语音波形。在语音库相对较小的情况下,这类方法可能得到较稳定的合成效果. 其缺点是往往存在声学特征参数“过平滑”问题,另外声码器也可能对音质造成损伤。

拼接合成指基于单元挑选和波形拼接的语音合成<sup>[3]</sup>. 其训练阶段与参数合成方式的基本相同,但在合成阶段通过模型计算代价来指导单元挑选,并采用动态规划算法选出最优单元序列,最后对选出的单元进行能量规整和波形拼接。拼接合成直接使用真实的语音片段,能最大限度保留语音音质. 缺点是一般需要较大音库,且无法保证领域外文本的合成效果.

传统的语音合成系统都是相对复杂的,例如: 前端需要较强的语言学背景,不同语言的语言学知识差异明显,通常需要特定领域的专家支持;后端的参数系统需要对语音的发声机理有一定了解,而传统参数系统建模时难以避免信息损失,限制了合成语音表现力的提升;同在后端的拼接系统对语音库要求较高,也常需人工介入指定挑选规则和参数<sup>[4]</sup>.

为改善这些问题,新的语音合成方式应运而生. 其中端到端合成便是一种非常重要的发展趋势. 在这种模式下,将文本或者注音字符输入系统,而系统则直接输出音频波形. 这降低了对语言学知识的要求,有利于表现更丰富的发音风格和韵律感,也可以相对方便地支持不同语种.

近年来语音合成发展迅猛,如谷歌的 Tacotron、Tacotron 2、WaveNet、Parallel WaveNet<sup>[5-8]</sup>, 百度的 DeepVoice、DeepVoice 2、ClariNet<sup>[9-11]</sup>, 英伟达的 WaveGlow<sup>[12]</sup>等. Tacotron 是第一个真正意义上的端到端语音合成系统,它允许输入文本或注音字符,输出线性谱,再经过声码器 Griffin-Lim 转换为波形. Tacotron 2 在 Tacotron 的基础上进行了模型简化,去掉了复杂的 CBHG (1-D Convolution Bank+Highway Network+Bidirectional GRU (Gated Recurrent Unit))结构, 使用了新颖的注意力机制 Location-Sensitive Attention, 提高了对齐稳定性. WaveNet 及其之后的 Parallel WaveNet 并非端到端系统,它们依赖其他模块对输入进行预处理,提供特征. 仿照 PixelRNN 图像生成方式, WaveNet 依据之前采样点来生成下一采样点,结构为带洞卷积<sup>[13]</sup>. 百度的 ClariNet 使用单高斯简化 Parallel WaveNet 的 KL(Kullback-Leibler) 目标函数,改进了蒸馏法算法,使得结构更简单稳定,并且通过 Bridge-net 连接了特征预测网络和 WaveNet,实现了端到端合成.

自 2017 年以来,端到端语音合成的研究进入了超高速发展时期,谷歌、百度和英伟达等研究机构不断推陈出新,在合成速度、风格迁移、合成自然度方面精益求精。然而,根据

领域内文献资料, 端到端语音合成目前仅能合成英文, 未见成型的中文系统。相较英文, 中文语音合成存在一些难点, 例如: 汉字不表音; 中文存在大量多音字和变调现象; 中文发音韵律较英文发音更为复杂, 如儿化音等。

本文设计了一种中文语音合成方案, 基于 Tacotron 2 在以下几个方面进行了改进。

- (1) 针对汉字不表音、变调和多音字等问题, 添加预处理模块, 将中文转化为注音字符。
- (2) 使用中文音频语料预训练 Tacotron 2 的解码器, 之后进行微调, 显著减少拼音-中文音频对所需的训练数据量。
- (3) 使用多层感知机代替停止符(Stop Token)处的线性变换, 显著减少合成急促停顿现象。
- (4) 利用 Transformer 中的多头注意力(MultiHead Attention)改进 Tacotron 2 的位置敏感注意力(Locative Sensitive Attention)<sup>[14]</sup>, 使其能够捕获到更多语音信息, 提升合成音质。

## 1 相关工作

### 1.1 序列到序列生成模型

序列到序列的生成模型<sup>[15]</sup>将输入序列( $x_1, x_2, \dots, x_t$ )转化为输出序列( $y_1, y_2, \dots, y_T$ )。机器翻译通常先将源语言编码到隐空间, 然后再解码到目标语言, 有

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

$$S_t = f(S_{t-1}, y_{t-1}, c_t), \quad (2)$$

$$y_t = \text{decoder}(S_t, y_{t-1}, c_t), \quad (3)$$

其中,  $h$ 、 $S$  分别是编码器和解码器的隐状态,  $c$  是由注意力机制计算得来的上下文向量, 由编码器隐状态  $h$  加权进行计算, 即

$$c_i = \sum_{j=1}^t a_{ij} \times h_j, \quad (4)$$

其中, 权值  $a_{ij}$  越高, 表示第  $i$  个输出在第  $j$  个输入上分配的注意力越多, 受第  $j$  个输入的影响也就越大。 $a_{ij}$  通常被设计为隐状态  $h$  和  $S$  的函数。

在语音合成领域中, 一般先将文本编码到隐状态, 然后解码到梅尔频谱。

### 1.2 Tacotron 2

Tacotron 2 将英文作为输入, 直接从英语文本生成声音波形, 如图1所示: 输入文本经词嵌入后首先送入 3 层 CNN (Convolutional Neural Network) 以获取序列中的上下文信息, 接着进入双向 LSTM(Long Short-Term Memory) 组成的编码器。梅尔频谱(在训练阶段, 每次送入固定长度的真实频谱; 在推断阶段, 每次送入上一个时间步的输出)首先进入预处理网络, 预处理网络的输出与上一个时间步的上下文向量拼接送入 2 层 LSTM, LSTM 的输出被用作计算本时间步的上下文向量, 并且经线性映射后用来预测停止符和梅尔频谱。为了提取更为高维的特征, 用于预测梅尔频谱的 LSTM 输出被带残差的 5 层 CNN 组成的后处理网络提纯

优化, 最后输出梅尔频谱<sup>[6]</sup>.

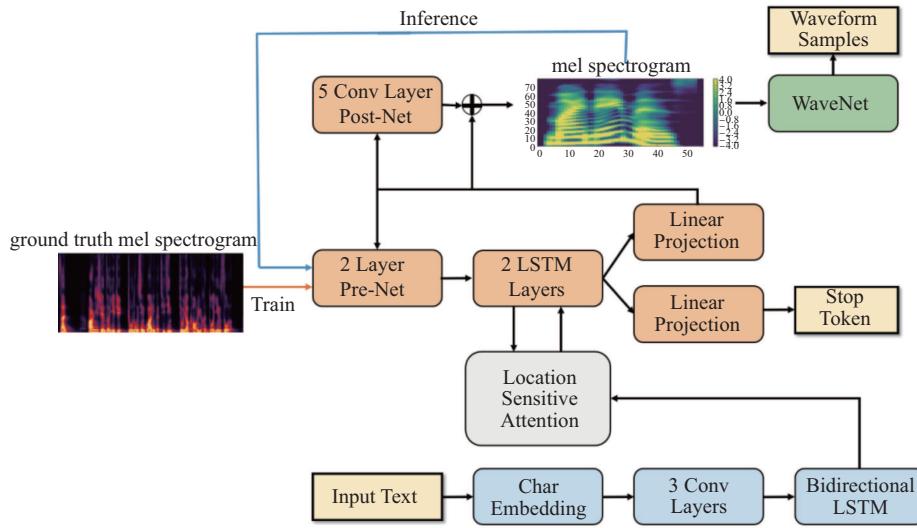


图 1 Tacotron 2 结构

Fig. 1 Architecture of Tacotron 2

### 1.3 Transformer

如图 2 所示, Transformer 是一种完全依赖注意力机制的端到端序列生成模型<sup>[14]</sup>, 在机器翻译领域显示出了其优异的性能。Transformer 模型的编码器由  $N$  个基本层堆叠而成: 每个基本层包含 2 个子层, 第一子层是 1 个 Attention, 第二子层是 1 个全连接前向神经网络。Transformer 模型的解码器也由  $N$  个基本层堆叠: 每个基本层除了与编码器相同的 2 个子层外, 还增加了 1 个掩码多头注意力子层, 所有子层都引入了残差边和 Layer Normalization。Transformer 的编、解码器都含有的多头注意力机制借鉴了 CNN 中多个卷积核的叠加, 实质上是将注意力机制独立执行几遍后拼接, 以更充分地抽取序列中的信息。

## 2 中文语音合成方案

与英文语音合成相比, 中文语音合成主要存在以下几个难点.

- (1) 无法直接使用中文作为文本输入, 需要添加文本–拼音/国际音标转换器, 并且要求在该预处理阶段解决中文变调和多音字问题.
- (2) 对语料要求较高, 需要保证说话人单一, 幅度变化小, 背景噪音小等. 相较 Tacotron 2 训练高音质的英文语音至少达 25 h<sup>[16]</sup>, 目前高质量中文语音合成语料较少.
- (3) 中文发音韵律变化较英文复杂.
- (4) 中文语音合成中, 往往发生语音生成急促停顿的现象, 尤其常见最后一个字无法正常发音的问题.

针对上述问题, 本文提出了一个基于 Tacotron 2 的中文端到端语音合成方案, 如图 3 所示, 并就文本–拼音转换器(Text to Phoneme)、预训练模块、注意力机制、停止符预测及后

处理等进行了特殊设计.

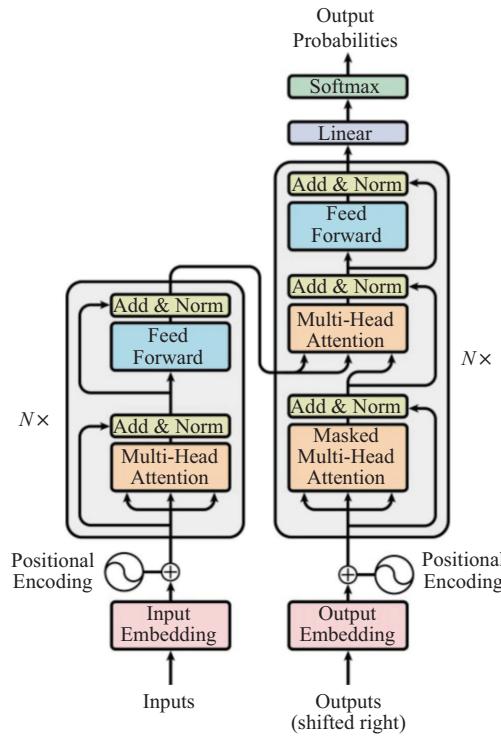


图2 Transformer 系统结构

Fig. 2 System architecture of transformer

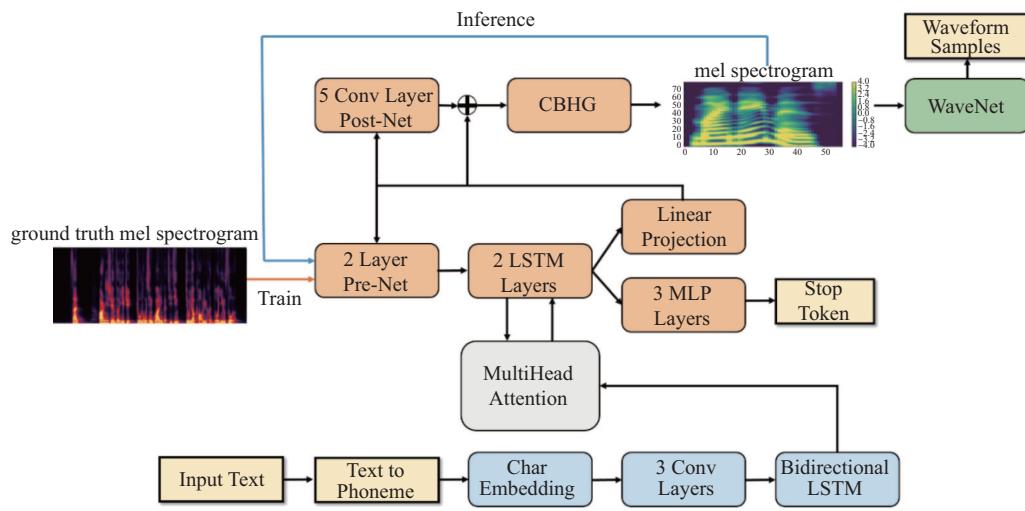


图3 中文端到端语音合成方案

Fig. 3 Proposed Chinese end-to-end speech synthesis scheme

## 2.1 文本-拼音转换器

不同于英文, 汉字不含发音信息, 可考虑先将中文转化为音素. 实验证明, 将中文转化为

拼音或国际音标, 合成后音质相差不大. 出于更加熟悉、便于纠错等原因, 本文最终采用拼音.

中文中存在变调现象, 如“第一”“十一”中的“一”读阴平, “一致”“一切”中读阳平, “一丝不苟”“一本万利”中读去声, 而“读一读”“看一看”则读轻声. 考虑到变调现象在中文中虽然存在, 但比例并不高, 本文采用规则匹配的方法解决. 对于多音字问题, 本文首先对输入文本进行中文分词, 然后利用词库对多音字进行正确注音. 通过上述方法, 基本可以正确地将中文文本转化为表音字符, 然后送入模型进行语音生成.

## 2.2 预训练

目前领域内的高质量拼音-音频合成语料稀少, 而 Tacotron 2 对语料的需要量却较大. 在本文中, 解码器使用中文音频进行初始化训练. 在预训练阶段, 解码器以教师指导模式预测下一个语音帧, 即以上一帧预测下一帧音频, 不需要对应的文本输入, 这要求解码器在帧级别学习声学自回归模型. 需要说明的是, 预训练阶段解码器仅依靠上一帧进行预测, 而微调阶段则需要基于解码器的额外输出进行推断, 这可能带来训练和推断的不匹配.

实验结果表明, 模型能有效学习语音中的声学信息, 并通过少量语料得到较好音质.

## 2.3 多头注意力机制

为了适用中文复杂的韵律变化, 本文将 Tacotron 2 中 Location-Sensitive Attention 扩展为 MultiHead Location-Sensitive Attention, 即

$$\text{head}_i = \text{Attention}(SW_i^s + HW_i^H + FW_i^F), \quad (5)$$

其中,  $S$  为当前时间步的解码器输出,  $H$  为编码器输出,  $F$  为累加注意力权重,  $W_i^s$ 、 $W_i^H$ 、 $W_i^F$  均为待训练参数, 每个子注意力模块权重不共享. 多头注意力输出为

$$\text{MultiHead}(S, H, F) = \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^o, \quad (6)$$

其中  $W^o$  表示待训练参数.

多头注意力将  $S$ 、 $H$ 、 $F$  通过参数矩阵映射再进行 Attention 运算, 然后把多个子注意力结果拼接起来. 类似于 CNN 中的多卷积核对一张图片提取特征的过程, 能够有效获取序列中的信息, 从而使解码器预测音频时, 字与字间的衔接以及整个句子的韵律变化更接近真实人声.

## 2.4 停止符预测和后处理网络

不同于 Tacotron 2 统一用线性变换预测梅尔频谱和停止符, 本文分别使用线性变换预测梅尔频谱, 用多层感知机(Multi-Layer Perceptron, MLP)预测停止符, 并且使用后处理网络优化重建梅尔频谱.

在中文合成过程中, 语音常常遇到戛然而止的现象, 影响语音流畅性. 类似问题在英文合成实践中也存在, 但并不明显. 这种停顿感主要是由于停止符预测的正负样本不平衡造成的. 本文通过将线性变换改为 3 层 MLP 并在二元交叉熵上加权(实验中将该权重设置为 6.0), 较好地解决了生成过程突然停顿的问题.

另外, 由于 Tacotron 2 的 WaveNet 生成较慢, 本文使用 Griffin-Lim 作为声码器<sup>[17]</sup>, 同时在原有的后处理网络添加 CBHG, 显著提高了音质.

### 3 实验结果与分析

本文通过实验验证了本文所提框架的有效性.

#### 3.1 训练步骤

使用 4 块英伟达 P100 训练模型, 利用 8 h 私有的拼音-音频语料和 50 h 中文音频作为训练数据集. 私有数据集的前后均保持 100 ms 静音间隔, 其中批处理规模(Batch Size)设置为 32, 过小的规模将造成训练不稳定并影响合成音质, 过大则容易引发内存溢出问题.

#### 3.2 文本-拼音转换器

Tacotron 2 直接将英文文本输入模型进行训练, 因为英文字母在单词中的发音变化较少, 如字母“a”的发音只有 [ei]、[a:] 和 [æ]. 但即便抛开中文汉字的多音字问题, 希望模型直接学习每个汉字的发音都是不现实的; 将汉字转化为注音字符, 如国际音标或拼音, 是较可行的思路. 然而如果在数据预处理过程中, 注音标注若出错, 合成结果必将失败. 本文通过规则匹配法基本解决了这类问题, 但该方法也存在局限性, 仍有少量(低于4%)汉字注音出现错误.

#### 3.3 Griffin-Lim 设置

由于 WaveNet 生成速度过慢的问题尚未解决, 本文选用 Griffin-Lim 作为模型的声码器, 迭代次数设置为 30. Tacotron 2 直接使用带残差的 5 层 CNN 作为后处理网络, 但其对梅尔频谱的优化不充分, 因此添加 CBHG 进一步提取特征以有效提升音质. 在实验中, 通过将原始的录音音频转化为梅尔频谱, 再使用 Griffin-Lim 转换回来, 发现有明显的音质损伤, 可以推断 Griffin-Lim 是影响音质的瓶颈. 另外为了进一步减少信息损失, 本文将梅尔频谱的输出维度由 80 改为 160.

#### 3.4 合成音频样例

本文提供了一些中文合成样例, 参见 <https://github.com/cnlinxi/tacotron2/tree/master/samples>. 这些样例由任意给定的中文文本通过文本-拼音转换器转化为拼音后, 输入已经训练好的模型合成. 模型训练利用 8 h 拼音-音频样本和 50 h 的音频样本, 训练步数为 15 万步, 每步耗时约 3.3 s.

#### 3.5 剥离分析

##### 3.5.1 预训练

当前公开的质量较高的中文语音合成功能为 THCHS-30<sup>[18]</sup>, 该数据集音频时长约为 30 h, 其对应的拼音标注较准确. 但是 THCHS-30 中有多个说话人, 男女声混杂, 背景噪音很大. 利用语料训练后合成的语音有的音频为男声, 有的为女声, 甚至同一句话一半为男声一半为女声, 并且合成后音质较差. 鉴于上述, 本文考虑使用 8 h 高质量私有语料进行预训练. 但 Tacotron 2 合成高质量语音通常要求较大的训练样本量, 因此需要一种减少训练样本需求的方法. 文献 [19] 提出使用预训练的词向量(英文)以减少 Tacotron 2 训练样本, 思路具有启发性. 考虑到汉字不表音, 本文曾考虑使用预训练的拼音词向量以增强信息, 但拼音预训练词向量非常罕见, 资源难以获得. 实验中发现, 通过对解码器使用单独的中文音频进行预训练, 也能获得较好的初始化效果. 特别地, 在预训练冻结编码器时, 解码器的输入端应给予轻微的扰动值, 以减小预训练和微调不匹配时带来的误差. 图 4 分别给出了 10 万步时使用解码器预训练和没有使用解码器预训练获得的梅尔频谱. 从图 4 可以看到, 相较前者, 后者锐利而清晰.

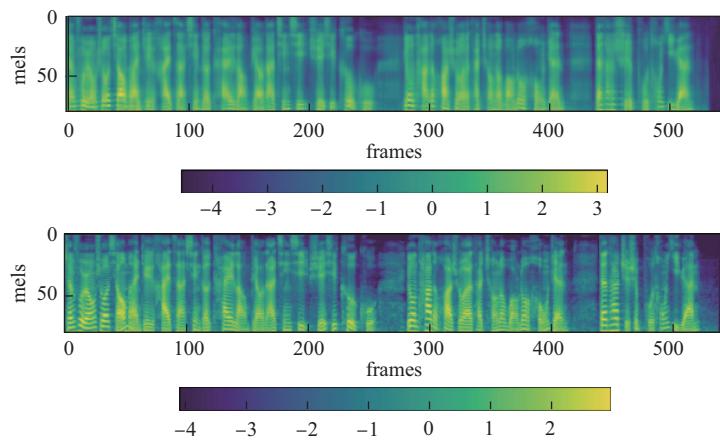


图 4 10 万步时预训练和未预训练获得的梅尔频谱

Fig. 4 100,000-step pre-trained and un-pretrained Mel spectrum

### 3.5.2 多头注意力机制

多头注意力机制能够对特定序列通过多个角度反复提取信息，并将各个子注意力模块的输出结果进行拼接，令生成语音时使用的信息更丰富，提升了合成语音音质。工程上可以使用梅尔倒谱距离(Mel Cepstral Distance, MCD)来评价音质，其值越小越好<sup>[20]</sup>。表 1 给出了使用 10 min 左右(213 句)验证集在 10 万步的模型上计算得到的 MCD。本文还对比了不同头数注意力机制对合成语音的影响，可以看到，头数的增加可能提高生成语音的质量。但同时也将使得训练速度变慢，内存占用增大，收敛速度减缓，因此未来存在优化的必要。

**表 1 原始 Tacotron 2、添加双头 (2-head) 和 4 头 (4-head) 的注意力机制之间的 MCD 比较**

Tab. 1 Comparison of MCD among original Tacotron 2 and improved ones (2-head, 4-head)  
with different attention mechanisms

	MCD
Tacotron 2-Base	20.03
2-head	18.1
4-head	17.26

### 3.6 与其他语音合成系统的比较

如表 2 所示，经 15 万步、4 头注意力训练得到的中文 Tacotron2，其 MCD 的值为 17.11，与文献 [19] 给出的 18.06 具有可比性。

**表 2 中文 Tacotron 2、英文 Tacotron 2 的 MCD 比较**

Tab. 2 Comparison of MCD between Chinese Tacotron 2 and English Tacotron 2

	MCD
中文 Tacotron 2	17.11
英文 Tacotron 2	18.06 <sup>[21]</sup>

文献 [19] 中英文 Tacotron 2 的评价印象分(Mean Opinion Score, MOS)，即人类主观评分为  $4.526 \pm 0.066$ ，优于文献 [2] 中拼接式语音合成系统的  $4.166 \pm 0.091$  和文献 [3] 中参数式语音合成系统的  $3.492 \pm 0.096$ 。

需要说明的是, 本文在实验中合成相同的 64 句话, 合成音频时长为 331.5 s, 耗时 366.11 s, 暂时无法满足实时要求.

## 4 总 结

本文设计并通过实验验证了一个基于 Tacotron 2 的中文 CNN 语音合成方案, 在语料有限的情况下, 可以实现端到端的较高质量中文语音合成. 梅尔频谱、梅尔倒谱距离等的实验对比结果表明了其有效性, 可较好地适应中文语音合成的要求; 就目前业内一般仅能端到端语音合成英文的局面, 是一个有益探索.

但本文方案目前尚存在一些问题, 如: 中文多音字辨识没有得到彻底解决; 合成语音中无法完全避免杂音, 仍存在少量不合理停顿现象; 对实时性的支持有待改善等. 今后可持续进行优化并开展较大规模人类主观评测.

## [参 考 文 献]

- [1] MOHAMMADI S H, KAIN A. An overview of voice conversion systems[J]. *Speech Communication*, 2017, 88: 65-82.
- [2] GONZALVO X, TAZARI S, CHAN C A, et al. Recent advances in Google real-time HMM-driven unit selection synthesizer[C]//Interspeech 2016. 2016: 2238-2242.
- [3] ZEN H, AGIOMYRGIANNAKIS Y, EGBERTS N, et al. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices[C]//Interspeech 2016. 2016: 2273-2277.
- [4] TAYLOR P. *Text-to-Speech Synthesis*[M]. Cambridge: Cambridge University Press, 2009.
- [5] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv preprint arXiv:1703.10135, 2017.
- [6] 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). IEEE, 2018: 4779-4783.
- [7] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.
- [8] OORD A, LI Y, BABUSCHKIN I, et al. Parallel WaveNet: Fast high-fidelity speech synthesis[J]. arXiv preprint arXiv:1711.10433, 2017.
- [9] ARIK S O, Chrzanowski M, Coates A, et al. Deep voice: Real-time neural text-to-speech[J]. arXiv preprint arXiv:1702.07825, 2017.
- [10] ARIK S, DIAMOS G, GIBIANSKY A, et al. Deep voice 2: Multi-speaker neural text-to-speech[J]. arXiv preprint arXiv:1705.08947, 2017.
- [11] PING W, PENG K, CHEN J. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech[J]. arXiv preprint arXiv:1807.07281, 2018.
- [12] PRENGER R, VALLE R, CATANZARO B. WaveGlow: A Flow-based Generative Network for Speech Synthesis[J]. arXiv preprint arXiv:1811.00002, 2018.
- [13] OORD A, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[J]. arXiv preprint arXiv:1601.06759, 2016.
- [14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//31st Annual Conference on Neural Information Processing Systems. NIPS, 2017: 5998-6008.
- [15] SUTSKEVER I, VINYALS O, Le Q V. Sequence to sequence learning with neural networks[C]//28th Annual Conference on Neural Information Processing Systems. NIPS, 2014: 3104-3112.
- [16] FREEMAN P, VILLEGRAS E, KAMALU J. Storytime-end to end neural networks for audiobooks[R/OL].[2018-08-28]. [http://web.stanford.edu/class/cs224s/reports/Pierce\\_Freeman.pdf](http://web.stanford.edu/class/cs224s/reports/Pierce_Freeman.pdf).
- [17] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform[J]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984, 32(2): 236-243.
- [18] WANG D, ZHANG X W. Thchs-30: A free chinese speech corpus[J]. arXiv preprint arXiv:1512.01882, 2015.
- [19] CHUNG Y A, WANG Y, HSU W N, et al. Semi-supervised training for improving data efficiency in end-to-end speech synthesis[J]. arXiv preprint arXiv:1808.10128, 2018.
- [20] KUBICHEK R. Mel-cepstral distance measure for objective speech quality assessment[C]//Communications, Computers and Signal Processing, IEEE Pacific Rim Conference on. IEEE, 1993: 125-128.

(责任编辑: 李 艺)