DOI：10.13890/j.issn.1000-128X.2023.06.015　　文章编号：1000-128X(2023)06-0122-07

轨道交通车载端到端语音合成

罗　潇，刘　悦
（中车株洲电力机车研究所有限公司，湖南 株洲　412001）

摘　要：高自然度的语音合成是车载人机交互进入高级智能化的基本要求之一。现阶段的轨道交通领域仍在广泛使用传统的低自然度语音合成算法，这与高速发展的智能化人机交互技术脱节。相比之下，端到端的深度学习语音合成算法凭借其近乎媲美人声的自然度已经成为各领域语音合成的主流算法。文章介绍了一种适用于离线轨道交通车载环境的端到端深度学习语音合成算法，该算法的主观意见评分达到4.18，并且在车载嵌入式硬件平台英伟达Xavier上的实时率达到0.52。试验证明，该算法不仅在自然度等主观性能上远超传统语音合成算法，同时也具备在轨道交通离线车载环境下的工程实用性。

关键词：轨道交通；车载人机交互；智能化；深度学习；端到端语音合成
中图分类号：U270.38^{+}2　　**文献标识码：**A

An end-to-end text-to-speech system for vehicle-mounted devices

LUO Xiao, LIU Yue

(CRRC Zhuzhou Institute Co., Ltd., Zhuzhou, Hunan 412001, China)

Abstract: High-naturalness text-to-speech is one of the basic requirements for advanced intelligence in vehicle-mounted human-machine interaction. Currently, in the rail transit field, there is widespread use of traditional low-naturalness text-to-speech algorithms, which are out of touch with the rapidly developing intelligent human-machine interaction technology. In contrast, end-to-end deep learning-based text-to-speech algorithms, with their nearly human-like naturalness, have become dominant in various fields of text-to-speech. This paper introduced an end-to-end deep learning-based text-to-speech algorithm suitable for offline railway vehicle environments. The mean opinion score of this algorithm reached 4.18, and the real-time rate on the vehicle-mounted embedded hardware platform NVIDIA Xavier reached 0.52. Experiments show that this algorithm not only outperforms traditional text-to-speech algorithms in terms of subjective performance such as naturalness, but also possesses engineering practicality in the offline vehicle environment of railway transportation.

Keywords: rail transit; vehicle-mounted human-machine interaction; intelligent; deep learning; end-to-end text-to-speech

0　引言

车载语音交互因其免接触的便捷性和无需视线注意力转移的安全特性，已经成为车载人机交互的一种主流方式和重要研究方向，轨道交通领域也早已开展语音交互相关的研究[1-2]，而语音合成在语音交互中扮演了闭环交互流程的重要角色。基于计算机的语音合成研究可追溯到20世纪下半叶[1]，在过去的几十年中，语音合成算法可以分为五大类：发音器官参数合成法[2]、共振峰合成法[3]、波形拼接合成法[4]、统计参数合成法[5]，以及近年随着深度学习技术迅速发展而出现的神经网络合成法[6-10]。一般而言，将前4种语音合成算法统称为传统算法，以与神经网络合成法区分。在传统算法中，发音器官参数合成法通过模拟人发音器官的发声过程来合成语音，然而在实践中很难准确地对发音行为进行建模，这种算法合成的语音质量并不高。共振峰合成法是一种基于专家规则的合成算法，

通信作者：罗　潇（1993—），男，硕士，工程师，主要从事语音信号处理工作；E-mail: shawl336@163.com
引用格式：罗潇, 刘悦. 轨道交通车载端到端语音合成[J]. 机车电传动, 2023(6): 122-128.
Citation：LUO Xiao, LIU Yue. An end-to-end text-to-speech system for vehicle-mounted devices[J]. Electric drive for locomotives,2023(6): 122-128.

包含 1 个用于生成波形的变参数声学模型和 1 个波形叠加模块，尽管共振峰合成法可以产生比较易懂的语音，但是仍然和自然的人声有明显的差别[1]。相比之下，波形拼接合成法通常以音素或音节为基元，并将预录制的基元时域波形与其相关的文本特征和声学特征预先储备于数据库中，然后依据这些相关特征的拼接得分找到最佳基元链接路径，并最终拼接对应的波形得到语音。因此，波形拼接法具有真实的音色和很高的可懂度，然而基元数据库需要占用大量内存，并且拼接而来的声音往往缺乏自然度和情感是该方法不可避免的缺点[1]。统计参数合成法很好地解决了拼接合成法对大量预录制发音音节的依赖，只需要少量预训练好的模型参数。它将语音合成分成了 3 个串联的主要模块，即前端文本特征提取、声学模型和声码器模型。其中，文本特征提取同样依据专家规则将自然语言文本转换为表征一系列语言学特征（音素位置、词性等）的符号文本；接着，声学模型再依据语言学特征预测出基频对数、倒谱系数等可以被声码器转换为声音波形的声学特征参数。可接受的自然度、灵活可调的模型参数和低内存要求是统计参数合成法的优点，但是仍然依赖有限的专家知识来制定特征提取规则，这使得其合成的语音可懂度较低，往往伴随有嗡鸣声、噪声以及明显区别于真人的机器感[1]。语音合成在拟人自然度上难以取得突破性的进展，直到深度学习技术被广泛应用，这种技术的普及在许多领域超越并取代了传统方法[6-7]，神经网络语音合成法也是如此。借助于深度学习网络强大的函数拟合能力和隐含特征自学习能力，神经网络合成法不需要单独地对声音的各种参数建模，仅需 1 个深度学习神经网络直接将文本转换为音频[8-9]或者先用 1 个声学神经网络，再配合 1 个声码器神经网络将声学特征转换为声音[10-13]，这种端到端的模式（即省去了人为提取特征的步骤，先预测声学参数再配合声码器的“两步走”合成模式）打破了传统算法中利用有限的专家知识来制定特征提取规则的限制，取得了前所未有的媲美人声的语音自然度和可懂度[8]。但需要明确的是，神经网络的推理过程往往伴随着大量的浮点计算，所以通常部署在云端，并不适合算力资源受限的嵌入式环境，比如车载嵌入式设备。然而，在轨道交通领域，公网的云端语音合成部署方案存在不可妥协的安全风险，私有化云端语音合成部署的成本又过高。目前，轨道交通主流车载语音合成算法采用的还是传统的统计参数语音合成算法。这些传统的算法尽管稳定，但是其自然度和可懂度都难以匹配迅速发展的深度学习技术带来的用户体验，也体现不出智能语音人机交互的“智能”感。基于此背景环境，本文将介绍一种 Tacotron 2[12]神经网络声学模型结合 LPCnet[14]神经网络声码器的全神经网络端到端语音合成系统，采用纯 CPU 运行，该系统不仅自然度和可懂度都远超传统统计参数方法，而且支持在车载嵌入式设备英伟达 Xavier 上的实时离线语音合成。

1　端到端神经网络语音合成系统

本文所提的端到端语音合成系统主要由 1 个神经网络声学模型串联 1 个神经网络声码器组成，其简要工作流程如图 1 所示。

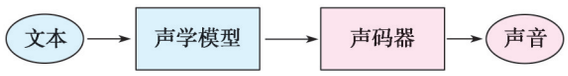

图 1　语音合成系统流程

Fig. 1　Process of the text-to-speech system

1.1　Tacotron 2 声学模型

神经网络在语音合成声学模型中的应用经历了从组件替换到全部取代传统模型的历程。文献[15]首先提出用最基础的全连接神经网络替换统计参数语音合成系统中的决策树模型，以提升对语言学特征到声学参数映射的函数拟合能力和解决决策树切分小样本空间导致的低泛化性。文献[16]则进一步加入长短期记忆网络[17]以替换隐马尔可夫模型，进而提升输入样本序列上下文之间的联系。文献[18-19]将统计参数语音合成算法中的传统模型全部替换为深度学习模型，但是仍然遵循着 3 模块的设计思路。Tacotron[11] 声学模型是第一个端到端的神经网络声学模型，它直接从音素序列预测线性声谱图，极大简化了前端特征提取和预测声学特征，类似的模型还有 Tacotron 2[12] 声学模型、DeepVoice 3[10] 声学模型和 FastSpeech 1/2[20-21] 声学模型等。这些后继的端到端模型主要在声音质量、合成速度和一步式合成等方面改进神经网络合成法。其中，Tacotron 2 声学模型的提出给神经网络合成法的工业化应用提供了充分的信心，而且该模型是在自然度和实时率综合表现最优的声学模型之一。在本文所设计的车载语音合成系统中，采用一种改进的 Tacotron 2 声学模型作为车载语音合成系统的声学模型。

Tacotron 2 模型是 Tacotron 模型的改进版本，属于典型的“编码-注意力-解码”声学模型。它能将输入文本映射为对应的声谱参数，比如梅尔谱倒谱参数、巴克倒谱参数（Bark Frequency cepstrum coefficients, BFCC），可根据后接声码器而定，其模型结构如图 2 所示。

图 2 中“浅蓝色”的部分为编码网络，它将输入文本的每个符号映射为对应的向量，这一过程一般通过 1 位有效编码与向量矩阵相乘的方法查表获取。3 层串联的卷积用来提取输入文本的词间相关特征，接着将最后一层卷积的输出送入 1 个单层双向长短期记忆网络，后者会循序渐进地为每一个输入的符号向量生成新的带有上下文环境的深度表示向量，即特征编码。位置

敏感注意力机制[22]会根据这些编码学习上下文特征，并参与到解码网络（图2“橙色”部分）的每一步，允许后者在解码时全局地参考这些特征编码。其数学模型为：

$$\varepsilon_{ij}=\boldsymbol{v}^{\mathrm{T}}\tanh\left(\boldsymbol{W}\cdot\boldsymbol{s}_i+\boldsymbol{V}\cdot h_j+\boldsymbol{U}\cdot f_{ij}+\boldsymbol{b}\right) \tag{1}$$

$$\boldsymbol{f}_i=\boldsymbol{F}*\boldsymbol{\alpha}_{i-1} \tag{2}$$

$$\boldsymbol{\alpha}_i=\mathrm{soft\,max}\left(\boldsymbol{\varepsilon}_i\right) \tag{3}$$

$$\boldsymbol{c}_i=\sum_j\left(\alpha_{ij}\cdot\boldsymbol{h}_j\right) \tag{4}$$

式中：j,i分别为编码和解码的步数索引；ε_{ij}为注意力能量；$\boldsymbol{v},\boldsymbol{b}$为权重向量，随着模型训练而自动更新；$\boldsymbol{W},\boldsymbol{V},\boldsymbol{U},\boldsymbol{F}$均为随机初始化矩阵，随着模型训练而自动更新；$\boldsymbol{s}_i$为零初始化的解码网络状态向量；$\boldsymbol{h}_j$为编码特征向量，是上一步编码网络的输出；$\boldsymbol{f}_i$为位置敏感特征向量；$\boldsymbol{\alpha}_i$为解码步数为$i$时对应的注意力权重向量；$\boldsymbol{c}_i$为最后求得的上下文特征向量，它综合了每一步特征编码对当前解码的相关程度。

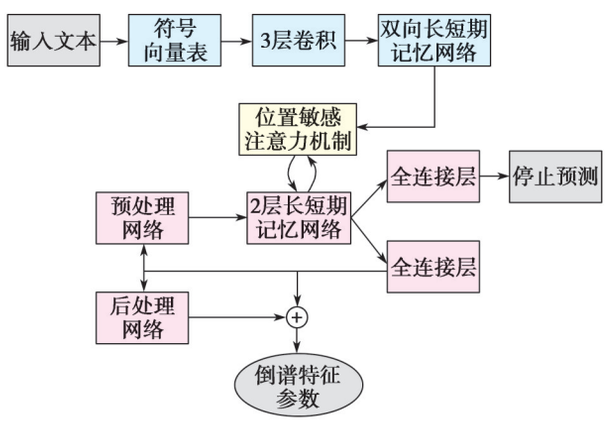

图2 Tacotron 2模型结构

Fig. 2 The architecture of Tacotron 2

解码网络的设计遵循自回归循环网络结构，它基于注意力机制生成的上下文特征向量预测声谱图参数序列，其核心组成为预处理网络、后处理网络、两层单向长短期记忆网络、2个全连接网络。预处理网络提供信息瓶颈的作用；后处理网络则预测真实声谱图参数与计算所得之间的残差，以提高预测质量；长短期记忆网络的每一步输出都平行地通过一个全连接网络得到每一帧的声谱图参数，并和另一个全连接网络判断是否该停止。

1.2 LPCNet声码器

直接预测声谱图参数的神经网络声学模型不能与传统基于信号处理的声码器兼容，一般需要配合另一个神经网络声码器将其预测得到的声谱图转换为原始声音波形。Wavenet声码器是第一个在合成声音质量上远超传统声码器的神经网络声码器，但是其依赖所有前序样本的自回归预测结构也使得其合成速度非常缓慢，无法应用在实时场景。因此，对Wavenet声码器提速也成为了一个重要的研究方向，Wavernn声码器[23]用门控循环网络简化了Wavenet声码器的自回归结构，并引入稀疏矩阵等加速技术，将单显卡合成速度提升了4倍。LPCNet声码器则是在Wavernn声码器的基础上继续融合了信号处理思想的一种更高效的神经网络声码器。

LPCNet声码器的结构如图3所示，其模型由3个主要部分组成，分别是帧预测网络、线性预测编码计算模块和样本预测网络。模型的输入特征为前文所提声学模型输出的倒谱特征参数，它包括每帧的BFCC和基音周期，以及基音周期的自相关矩阵。帧预测网络由2个残差卷积网络和2个全连接层组成，负责预测帧级别的声音特征。线性预测编码计算模块是1个非网络计算模块，它借鉴信号处理中的全极点滤波编码器来模拟发声过程，其基本原理是根据前面已知的M个样本的线性组合来预测新的样本，其计算公式为

$$p_t=\sum_{k=1}^{M}\left(a_k d_{t-k}\right) \tag{5}$$

式中：t为时间索引；p_t为线性预测编码的输出；d_{t-k}为前序的声音样本点；a_k为线性预测系数；M为一个超参数，它决定了滤波器的极点个数。

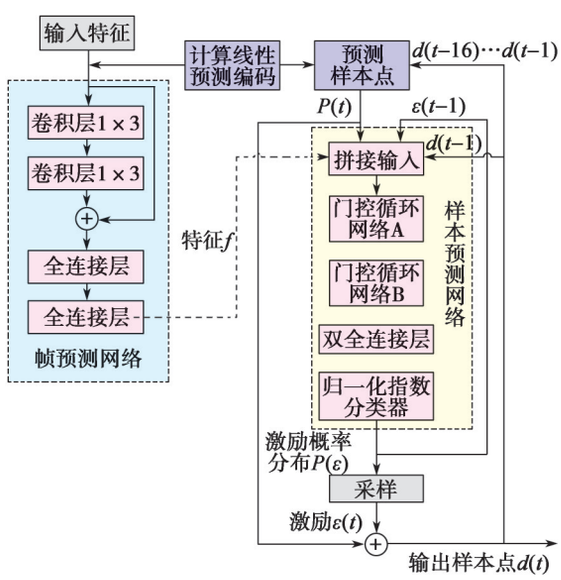

图3 LPCNet模型结构

Fig. 3 The architecture of LPCNet

线性编码系数的计算过程为先将BFCC转换为线性谱能量密度，再通过逆傅里叶变化得到信号的自相关矩阵，最后利用莱文森-杜宾算法从自相关矩阵中递归求得线性预测系数向量，即式(5)中的M个a。

事实上，由于BFCC是线性谱通过非线性滤波器组和离散余弦变化等一系列计算得到的有损特征表示，直接从BFCC计算得到的LPC系数相比从原始信号计算是不准确的，不过这一损失可以通过样本预测网络来学习弥补。样本预测网络由2个门控循环网络、2个平行的全连接层和1个归一化指数分类器依序组成，其

输入包括帧预测网络的输出f（在预测每帧内的所有样本点时为定值）、当前时刻t的初始预测样本点p、上一时刻$t-1$的输出样本点d、上一时刻$t-1$的样本预测网络输出ε。其中，ε的取值范围为1个256类的离散空间，可以视为信号的激励，时刻t的最终声音样本预测为

$$d_t=p_t+\varepsilon_t \tag{6}$$

2　模型优化与推理

原始的Tacotron 2声学模型并非完美，其性能和稳定性都存在优化和提升的空间，因此可在原始的模型结构或者模型推理上进行优化。

2.1　高斯混合模型注意力对齐机制

原始的Tacotron 2声学模型中采用的是位置敏感注意力对齐，这种注意力对齐机制在计算每一步的注意力权重时考虑了输入的内容和上一步的权重。然而，位置敏感注意力最初是针对语音识别提出来的，它并没有考虑语音合成中注意力权重的单调递增特性，比如，在处理包含重复内容的输入文本时，基于内容的部分[即式(1)中的$\boldsymbol{s}_i$和$\boldsymbol{h}_j$]）可能会受之前出现的相同内容影响，导致注意力往回看[24]。相比之下，高斯混合模型注意力作为一种更简单的相对位置注意力机制，能在一定条件下保证对齐向量的单调性。文献[24]给出了高斯混合模型注意力对齐的通用格式，如下所示：

$$\alpha_{ij}=\sum_{k=1}^{K}\left[\frac{w_{ik}}{Z_{ik}}\cdot \mathrm{e}^{-\frac{(j-\mu_{ik})^2}{2\sigma_{ik}^2}}\right] \tag{7}$$

$$\boldsymbol{\mu}_i=\boldsymbol{\mu}_{i-1}+\Delta_i \tag{8}$$

式中：K为高斯混合模型的子模型个数；k为子模型的索引；Z为归一化分母，本文取其为一个常量1；$\boldsymbol{w}_i,\Delta_i,\boldsymbol{\sigma}_i$均为$k$维向量，每个维度对应$k$个子高斯模型，它们由前一步的解码输出送入一个全连接神经网络预测得到；$\boldsymbol{\mu}_i$为1个k维向量，由k个子高斯模型的均值组成。

明显地，当Δ_i总为正数的时候，该高斯混合模型的均值随着解码步数i单调递增，并且高斯模型的分布特点保证了注意力权重的位置相对性（每个位置的权重与最大权重位置的相对距离有关，相对距离越远权重越小），这符合输入文本与输出音频自有的对齐特点。

2.2　流式语音合成

试验阶段的语音合成都是采用非流式的语音合成方法，即需要等待算法将所有的音频合成之后再播放，这种等待的代价在实际应用部署中是不可取的。Tacotron 2声学模型自循环式的解码方法先天适用流式语音合成，即每解码生成一定帧数的输出便送入LPCNet生成声音并立马播放。一般地，只要解码的时间小于合成声音的长度的0.9倍即可保证流程语音合成的连贯性。采用以10（5×2）帧为单位的流式解码，其中“5”是流式合成的解码间隔步数（动态可调），“2”是声学模型每步解码输出的声音帧数（在训练模型时固化），即声学模型的每步解码输出对应2帧声音，而每解码5步就会调用声码器将这10帧合成声音波形。

3　对比试验

针对轨道交通嵌入式智能计算平台英伟达Xavier，为了验证提出的语音合成系统在嵌入式设备上的可用性和发音自然度，与基于隐马尔可夫的主流传统语音合成算法（HTS）在流式合成实时率和主观平均意见评分两方面进行了对比。实时率是流式嵌入端语音合成的关键指标，实时率小于1（即计算花费的时间与合成语音的时间长度之比）才能达到部署的基本要求；另一个常用的评价指标是主观平均意见得分，人为地根据多个对比语音片段进行主观打分，最高为5分，最低为1分，打分间隔为0.5分，最后平均值取95%置信区间。

3.1　模型超参数设定

Tacotron 2声学模型的单个样本输入L_1×640的矩阵，其中L_1为文本长度，640维包含了512的词向量和128的音调描述向量；2层预处理网络的大小分别为512和256；编码网络输出、循环网络解码输出和注意力向量均为512维，后处理网络作为一个锦上添花的网络在实际实现中被略去。Tacotron 2声学模型的目标输出为10 ms帧移和25 ms帧长的BFCC。

HTS的输入以音素为单位，基于开源HTS标签并针对中文特点改进得到91个问题，即HTS的单个样本输入为L_2×91，L_2为音素级别的样本长度。对应HTS的“梅尔谱–线性谱对”（MGC-LSP）声码器，HTS输出采用的是Alpha为“0”，Gamma为“-1”的16阶梅尔通用系数（等价于线性预测编码系数）。

试验合成的音频均为16 000采样率，16位深度，并且所有对比算法共享包括文本清理、正则化、分词、词性标注、字音转换、韵律（停顿）预测等前端预处理过程。

3.2　注意力对齐机制对比

在保持数据集、其余参数和模型不变的前提下，分别采用8个高斯模型组成的高斯混合模型注意力和512维隐变量的位置敏感注意力进行训练，并对这2种注意力机制的对齐收敛速度和对齐效果进行了对比，对比效果如图4~图6所示。

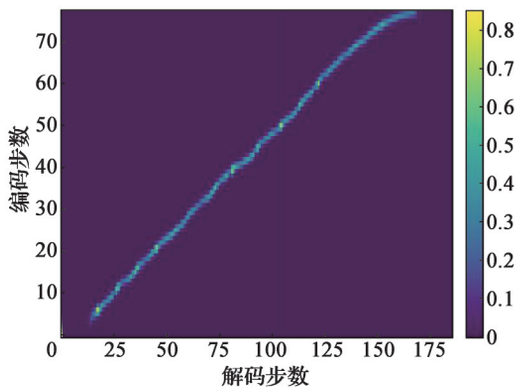

图4　高斯混合模型注意力对齐(第200 000步对齐矩阵)

Fig. 4　GMM alignment (at the 200 000th step alignment matrix)

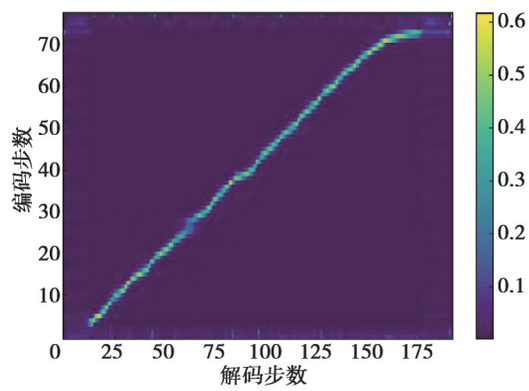

图5　位置敏感注意力对齐(第200 000步对齐矩阵)

Fig.5　Location-sensitive alignment (at the 200 000th step alignment matrix)

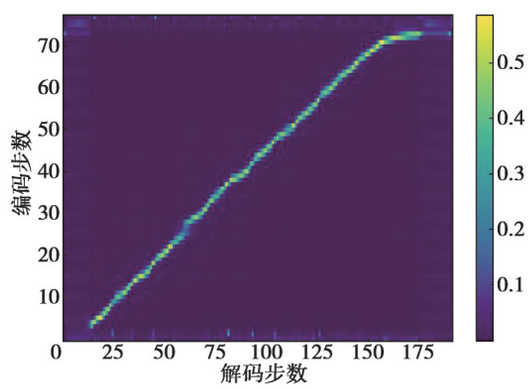

图6　位置敏感注意力对齐(第340 000步对齐矩阵)

Fig. 6　Location-sensitive alignment (at the 340 000th step alignment matrix)

由图4~6可知，高斯混合模型注意力机制在200 000步时已经明显收敛（实际在第80 000步就已经开始逐渐收敛），即注意力线条单调递增且不存在明显的模糊，而位置敏感注意力机制则在解码的约第60步和结尾处仍然存在模糊，即便继续训练到第340 000步时仍然无法收敛，这种模糊现象反应到合成的音频上就会出现吐词不清晰、可懂度差。对齐矩阵的对比证明了高斯混合模型注意力机制在实践中具有更快的收敛速度和更稳定的对齐能力。

3.3　平均主观意见评分对比

平均主观意见评分能在一定程度上量化地评价语音的自然度、可懂度和相似度等主观感受。在试验过程中，随机邀请了30位个体对匿名准备的音频进行主观打分，分别包括了真人原始音频、基于传统隐马尔可夫加高斯混合模型合成的音频（用HTS表示）和基于端到端深度学习模型合成的音频。其中，端到端模型还被细分为基于“原始Tacotron 2声学模型+LPCNet声码器”（Loc-Taco + LPCNet）的方案和本文提出的采用高斯混合模型注意力机制的“Tacotron 2声学模型+LPCNet声码器”（GMM-Taco + LPCNet）的方案，对比结果如表1所示。

表1　平均主观意见评分对比

Table 1　Comparison of mean opinion scores

语音合成模型	平均主观意见评分
HTS	3.68±0.12
Loc-Taco + LPCNet	3.72±0.11
GMM-Taco + LPCNet	4.18±0.08
原始音频	4.76±0.09

尽管在文献[12]中原始Tacontron 2声学模型在不考虑算力限制的情况下可以合成媲美人声的语音，但试验发现原始的Tacotron 2声学模型在实践中的表现不理想，其合成的语音常出现个别字发音模糊，导致其平均主观意见评分远低于预期。相比之下，采用高斯混合模型注意力机制的端到端模型合成的语音稳定且未出现发音模糊等现象，其评分自然也远高于前者。虽然2种对齐机制下的端到端模型评分差距不能直接体现两者在合成语音自然度等性能方面的差异，但是可以进一步地反映出高斯混合模型注意力机制的高稳定性。另外，本文提及的端到端语音合成系统相较于隐马尔可夫加高斯混合模型合成系统在评分上提升了一个打分量级，这说明其合成的音频在自然度、可懂度等主观感受方面有质的提高。不过，受限于终端嵌入式设备的算力，嵌入式的端到端深度学习语音合成系统还没能完全媲美真人原始语音，在主观听感上还存在进步的空间。

3.4　合成实时率测试

合成实时率定义为算法模型合成单位长度音频的实际时间开销与音频时间长度的比值，即合成每秒声音的计算开销。对实时率的要求一直是制约深度学习技术在终端嵌入式设备上推理的一个主要指标。本文用100条包含不同长度的句子分别对本文所提的端到端深度学习语音合成系统和基于隐马尔可夫加高斯混合模型的语音合成系统进行了合成实时率测试（纯CPU运行），并取其平均值作为最终的实时率，其结果如表2所示。

以合成1 s音频为参照基准，2种端到端合成模型需要分别计算约521 ms和602 ms，即0.521和0.602的实时率，是实时语音的约1.9倍和约1.7倍。基于高斯

混合模型注意力机制的端到端模型在推理时比位置敏感注意力机制的端到端模型计算更高效，这主要归功于前者更少的模型参数（约725万对比约751万），但两者均满足实时语音合成的要求。理论上，只要实时率小于1就是实时语音合成，所以在实际应用中，根据合成声音的速度快慢，动态地降低实时率，使其基本保持在0.8~0.9，以尽可能地降低处理器的负载。

表2　实时率对比

Table 2　Comparison of real-time rate

语音合成模型	实时率/倍数
HTS	0.027/37.04倍
Loc-Taco + LPCNet	0.602/1.66倍
GMM-Taco + LPCNet	0.521/1.92倍
原始音频	1/1倍

需要注意的是，基于隐马尔可夫和高斯混合模型的传统语音合成系统（HTS）实时率相当之高，达到实时语音的37倍。这确实是传统方法的优势，另一方面也是由于传统方法受限于对复杂生物发声系统的完全认知，建模相对简单的缘故。端到端模型利用深度学习技术去黑盒学习生物发声原理，虽然模型更加复杂和耗时，但是也带来了更好的合成自然度和可懂度等。

4　结语

基于深度学习神经网络的端到端算法是目前综合自然度、可懂度等主观评价最优的语音合成算法，本文介绍了一种适用于轨道交通嵌入式环境的端到端神经网络语音合成系统。该合成系统由1个优化了对齐机制的Tacotron 2声学模型和1个高效的LPCNet网络声码器组成。平均主观意见评分对比表明，该端到端语音合成系统在自然度和可懂度等主观评价上相较于传统的车载统计参数语音合成系统提升了一个评分量级，具有明显的优越性。同时，针对英伟达Xavier车载嵌入式设备，本文所提的端到端语音合成可以以1.92倍的速度实时地合成语音流，具备嵌入式语音合成系统部署应用的硬性条件。

参考文献：

[1] 刘悦，林军，游俊．语音识别技术在车载领域的应用及发展[J]．控制与信息技术，2019(2):1-6.
LIU Yue, LIN Jun, YOU Jun. Application and development of automatic speech recognition in vehicle field[J]. Control and information technology, 2019(2):1-6.

[2] 刘悦，林军，罗潇，等．基于时延神经网络的语音识别算法及其在轨道交通领域的应用研究[J]．控制与信息技术，2022(4):11-16.
LIU Yue, LIN Jun, LUO Xiao, et al. Research of speech recognition algorithm based on time delay neural network and its application in rail transit[J].Control and information technology, 2022(4):11-16.

[3] SEEVIOUR P, HOLMES J, JUDD M. Automatic generation of control signals for a parallel formant speech synthesizer [C]//IEEE. IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia: IEEE, 1976: 690-693.

[4] OLIVE J. Rule synthesis of speech from dyadic units[C]// IEEE. IEEE International Conference on Acoustics, Speech, and Signal Processing. Hartford: IEEE, 1977: 568-570.

[5] YOSHIMURA T, TOKUDA K, MASUKO T, et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis[C]//DBLP. Sixth European Conference on Speech Communication and Technology. Budapest: DBLP, 1999: 2347-2350.

[6] 罗江平, 喻熙倬, 曹经纬, 等. 基于深度学习与支持向量机的钢轨伤损智能识别系统[J]. 机车电传动, 2021(2): 100-107.
LUO Jiangping, YU Xizhuo, CAO Jingwei, et al. Intelligent rail flaw detection system based on deep learning and support vector machine[J]. Electric drive for locomotives, 2021 (2): 100-107.

[7] 贺德强, 江洲, 陈基永, 等. 基于深度卷积神经网络的铁路接触网鸟窝检测方法研究[J]. 机车电传动, 2019(4): 126-130.
HE Deqiang, JIANG Zhou, CHEN Jiyong, et al. Research on detection of bird nests in overhead catenary based on deep convolutional neural network[J]. Electric drive for locomotives, 2019(4): 126-130.

[8] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[C]//ISCA. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9). Sunnyvale: ISCA, 2016: 125.

[9] SOTELO J, MEHRI S, KUMAR K, et al. Char2Wav: end-to-end speech synthesis[C]//ICLR. International Conference on Learning Representations. Toulon: ICLR, 2017: 1-6.

[10] PING Wei, PENG Kainan, GIBIANSKY A, et al. Deep voice 3: scaling text-to-speech with convolutional sequence learning[C]//ICLR. 6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-16.

[11] WANG Yuxuan, SKERRY-RYAN R J, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[C]//ISCA. Interspeech 2017. Stockholm: ISCA, 2017: 4006-4010.

[12] SHEN J, PANG Ruoming, WEISS R J, et al. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions[C]//IEEE. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018: 4779-4783.

[13] LI Naihan, LIU Shujie, LIU Yanqing, et al. Neural speech synthesis with transformer network[C]//AAAI. The Thirty-

third AAAI Conference on Artificial Intelligence (AAAI-19). Hawaii: AAAI Press, 2019: 6706-6713.

[14] VALIN J M, SKOGLUND J. LPCNET: improving neural speech synthesis through linear prediction[C]//IEEE. ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2018: 5891-5895.

[15] ZEN Heiga, SENIOR A, SCHUSTER M. Statistical parametric speech synthesis using deep neural networks[C]//IEEE. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 7962-7966.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[17] FAN Yuchen, QIAN Yao, XIE Fenglong, et al. TTS synthesis with bidirectional LSTM based recurrent neural networks [C]//ISCA. Interspeech 2014. Singapore: ISCA, 2014: 1964-1968.

[18] ARIK S Ö, CHRZANOWSKI M, COATES A, et al. Deep voice: real-time neural text-to-speech[C]//JMLR. Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR, 2017: 195-204.

[19] ARIK S Ö, DIAMOS G, GIBIANSKY A, et al. Deep voice 2: multi-speaker neural text-to-speech[C]//Curran Associates Inc.. Proceedings of the 31st International Conference on Neural Information Processing Systems. California: Curran Associates Inc., 2017: 2966-2974.

[20] REN Yi, RUAN Yangjun, TAN Xu, et al. FastSpeech: fast, robust and controllable text to speech[C]//Curran Associates Inc.. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019: 3171-3180.

[21] REN Yi, HU Chenxu, TAN Xu, et al. FastSpeech 2: fast and high-quality end-to-end text to speech[C]//ICLR. International Conference on Learning Representations. Virtual: ICLR, 2021: 1-15.

[22] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]//MIT Press. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015: 577-585.

[23] KALCHBRENNER N, ELSEN E, SIMONYAN K, et al. Efficient neural audio synthesis[C]//PMLR. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018: 2410-2419.

[24] BATTENBERG E, SKERRY-RYAN R J, MARIOORYAD S, et al. Location-relative attention mechanisms for robust long-form speech synthesis[C]//IEEE. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 6194-6198.

收稿日期：2022-03-24
修回日期：2023-05-23