计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800127

基于交替训练及预训练的低资源泰语语音合成

蔡浩然[1]　杨　鉴[1]　杨　琳[1]　刘　聪[2]
1 云南大学信息学院　昆明 650504
2 科大讯飞股份有限公司人工智能研究院　合肥 230088
(chr164663553@163.com)

摘　要　泰语作为一种有数千万人口使用的语言,应用较为广泛,20世纪90年代末就有学者开展了泰语语音合成的研究。近年来,基于深度神经网络并利用大规模高质量“文本-音频”数据训练的端到端语音合成系统,已经能够合成出高质量的语音。目前,汉语、英语等通用语已拥有海量的语音合成数据库,然而泰语作为一种非通用语可获取的“文本-音频”数据库规模往往较小。在低资源条件下,以提高泰语语音合成质量为目标,选用端到端语音合成模型 Tacotron2 作为基线模型,研究交替训练方法以及预训练方法,研究不同文本嵌入方式对泰语语音合成效果的影响;然后从注意力对齐图和 MOS 评分两方面对文中设计的6种模型所合成的语音进行测评。实验结果表明,采用“元辅音嵌入+预训练+交替训练”方法的系统的语音合成质量最好,合成语音的 MOS 评分可达3.95分,明显优于基线系统的1.71分。

关键词:语音合成;泰语;低资源;交替训练;预训练
中图法分类号　TP391

Low-resource Thai Speech Synthesis Based on Alternate Training and Pre-training

CAI Haoran[1],YANG Jian[1],YANG Lin[1] and LIU Cong[2]
1 School of Information Science & Engineering,Yunnan University,Kunming 650504,China
2 AI Research Institute,iFLYTEK Co.,Ltd.,Hefei 230088,China

Abstract　As a language spoken by tens of millions of people,Thai is widely used. In the late 1990s,some scholars carried out research on Thai speech synthesis. In recent years,end-to-end speech synthesis systems based on deep neural networks and trained with large-scale high-quality “text-audio” data have been able to synthesize high-quality speech. At present,Chinese,English and other common languages have massive speech synthesis databases. However,the “text-audio” database available for Thai as a non-common language is often small in scale. Under the condition of low resources,this paper aims to improve the quality of Thai speech synthesis,selects the end-to-end speech synthesis model Tacotorn2 as the baseline model,studies the alternate training method and pre-training method,and studies the effect of different text embedding methods on the effect of Thai speech synthesis. Then,the speech synthesized by the six models designed in this paper is evaluated from the attention alignment map and the MOS score. Experimental results show that the system using the method of “vowel consonant embedding+pre-training+alternate training” has the best speech synthesis quality,and the MOS score of the synthesized speech can reach 3.95,which is significantly better than the baseline system's 1.71.

Keywords　Speech synthesis,Thai,Low resource,Alternate training,Pre-training

1　引言

语音合成即由文本生成自然语音,是一项极具挑战性的任务,也是近年来人工智能领域研究的热点。传统语音合成分为前端和后端两部分:前端主要对输入文本信息进行处理,包括分词、字素转音素、韵律预测等;后端主要将前端提取的信息利用某种技术转换为语音。传统的语音合成系统不仅结构复杂,而且需要具有较多的语言学相关知识储备才能完成数据准备工作,系统稳定性不高,存在误差累积,难以适用于各种复杂场景。随着深度神经网络在各领域的应用,一些专家学者开始研究将深度神经网络引入语音合成模型以提升模型的性能和简化模型的结构[1]。

近年来,以谷歌、百度为代表的一些机构相继推出了一系列的基于深度神经网络的端到端语音合成系统,如 Tacotron[2],Tacotron2[3],Fastspeech2[4] 以及 DeepVoice3[5] 等。这些端到端语音合成系统实现了真正意义上的从文本到语音的生成,并解决了传统语音合成模型的误差累积问题。谷歌的 Tacotron 是第一个较为成熟的端到端语音合成系统,它的骨干包含一个编码器、一个注意力机制解码器和一个后处理网络。编码器用于提取文本可靠的序列表示,解码器则学习文本与音频的对齐关系并生成合成语音的 Mel 谱。最终由后处理网络中的 Griffin Lim 重建算法生成最终的合成语音

基金项目:国家重点研发计划(2020AAA0107901)
This work was supported by the National Key Research and Development Program of China(2020AAA0107901).
通信作者:杨鉴(jianyang@ynu.edu.cn)

波形。其改进版本 Tacotron2 在此基础上优化了编码器和解码器的结构，并且增加了 Post-net 部分来对生成的 Mel 谱进行更精细的调整。使得系统的合成效果达到了较高的水平。

泰语是属于汉藏语系壮侗语族的一种语言，是泰国的官方语言，除泰国本土外，主要由分布在老挝、缅甸、越南西北、中国西南等地的傣泰民族使用。虽然泰语使用较为广泛，但其电子语料资源十分匮乏，导致近年来泰语的语音合成研究进展十分缓慢。正如文献[6]所介绍，泰语的语音合成研究自上世纪 90 年代末就已经广泛展开了。在 2003 年出现了基于单元选择方法以及大型语音语料库的泰语语音合成工作。文献[7]搭建了基于 HMM 的泰语语音合成系统。我们注意到，基于传统的统计参数模型或者 HMM 的泰语语音合成系统需要考虑如语境标签、停顿标签、协同发音等大量语音学和语言学的知识才能取得较好的效果[8]，并且基于 HMM 的泰语语音合成系统要想取得较好的效果对语料库的要求也较高[9]。因此本文提出基于端到端的模型的泰语语音合成系统，不仅是为了在构建泰语语音合成系统前简化对文本和语音进行处理的准备步骤，而且使得最终合成的泰语语音仍然能够与文本准确对应且自然度较高。

但是，目前以 Tacotron2 为代表的端到端语音合成系统的设计思路主要以英语的语音合成为最终目的，需要大量高质量的语音以及与之相对应的文本来训练模型。因此，这些端到端语音合成系统往往可以在英语、汉语等通用语上获得极佳的效果。相较于英语和汉语，由于缺乏大规模高质量的泰语语音文本对这一客观因素，直接把 Tacotron2 等端到端语音合成系统应用在泰语的语音合成上效果不佳。为此，本文设计并实现了一个基于 Tacotron2 的端到端泰语语音合成模型，采用交替训练的方法来提高 Tacotron2 系统的稳定性和加快模型的收敛，并且提出了一种针对泰语的预训练方法来提高系统的性能，在网络的前端部分提出并探讨了泰语的字符嵌入和元辅音嵌入的差异。

本文第 2 节介绍了本文构建的端到端语音合成系统，包括本文使用的模型以及对模型的改进；第 3 节介绍了实验方案和实验结果；最后总结全文。

2 模型与训练方法

2.1 Tacotron2 模型

本文采用了 Tacotron2 模型作为泰语语音合成模型的基线系统，Tacotron2 是典型的序列到序列模型，在目前语音合成研究中使用较为广泛，它主要由编码器和基于注意力机制的解码器组成。其模型如图 1 所示。编码器模块的作用是提取文本序列的特征以及上下文信息，它由 3 层一维卷积以及双向 LSTM 层组成。在字符嵌入层把文本输入序列的基本单元编码为 512 维的向量后，输入序列在编码器中经过 3 层卷积，每层卷积都包含 512 个 5×1 的卷积核，对输入序列进行上下文建模，最后通过双向 LSTM 层生成最终的编码特征。Tacotron2 的解码器是一个自回归网络，利用 Mel 谱的上一帧信息来预测下一帧信息，并预测停止符(Stop Token)，推断语音在何时结束。解码器的结构包括 2 层预处理网络，2 层 LSTM 以及 5 层后处理网络。在训练时，音频的真实 Mel 谱会与上一步的注意力上下文向量进行拼接，再和编码器输出的文本序列特征信息一起输入注意力机制，得到当前的注意力上下文向量和当前注意力权重，最终利用当前的注意力上下文向量和解码器的隐藏状态来预测当前帧的 Mel 谱以及停止标志。最后将预测到的完整的 Mel 谱通过 5 层卷积的后处理网络来得到更精细的 Mel 谱。

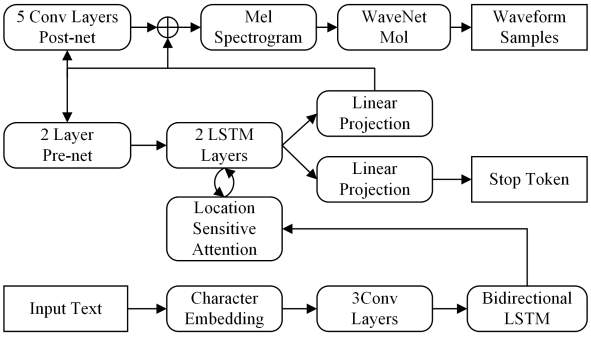


图 1 Tacotron2 模型图

Fig. 1 Diagram of Tacotron2 model

2.2 交替训练方法

Tacotron2 模型在训练阶段总是使用真实音频的上一帧来预测合成音频的下一帧，在推断过程中模型只能使用合成音频的上一帧来预测合成音频的下一帧，导致模型的训练和合成阶段存在一个暴露偏差的问题。这个差距在合成长句时尤其明显，表现为 Tacotron2 系统在推断时经常不能合成完整的句子，在合成音频中间出现大量静默段，并且会出现文本和音频不对应的情况。为了解决这一问题，本文在文献[10-11]的基础上提出了一种新的交替训练的方式。我们的做法是将整个训练过程分为 3 个阶段：第一阶段让 Tacotron2 系统采用固定概率使用合成音频的上一帧来预测合成音频的下一帧，随着训练步长的增加来到第二阶段；在第二阶段随着 epoch 的增加使这一概率逐渐增大；第三阶段保持第二阶段最后一个 epoch 的概率系数不发生改变。整个交替训练过程可表示为式(1)。

$$P(t)=\begin{cases}\dfrac{t_1}{n}, & 0<t\leqslant t_1\\ \dfrac{t}{n}, & t_1<t\leqslant t_2\\ \dfrac{t_2}{n}, & t_2<t\leqslant n\end{cases}\tag{1}$$

其中，$P(t)$为训练时使用合成语音帧的概率，显然，使用真实语音帧的概率为 $1-P(t)$，t 表示训练阶段的某一个 epoch，n 为训练阶段最后一个 epoch，t_1 为第一阶段最后一个 epoch，t_2 为第二阶段最后一个 epoch。通过多次实验得出，通常，n 为 500、t_1 取值范围为 50～100、t_2 取值范围为 200～250 时训练的结果优于基线系统。

之所以把这一概率设计成动态变化的，是我们发现在刚开始训练时，让系统以较低的概率使用合成音频的上一帧来预测合成音频的下一帧可以加快模型收敛，随后逐渐增大这一概率相当于对模型进行微调，最后一个阶段保持概率不变是为了使模型更加适应预测帧和自然帧的不同，并使系统达到稳定。第 3 节中设计的实验证实了所提出方法的有效性。

2.3 预训练方法

由于我们目前获取到的泰语语料相对较少，只采用泰语平行语料对系统进行训练的效果并不理想。受迁移学习的启发，本文提出一种预训练方法来缓解这一问题。

针对语音合成的迁移学习往往会应用于两种语法较为相似的语言中，对两种语言较为类似的发音单元进行总结和复用[12-13]。具体做法是先用汉语或英语的“文本-音频”训练得到一个汉语或英语语音合成系统，在目标语言在和源语言复用一些文本嵌入单元的基础上给模型加入目标语言独有的文本单元并共享网络权重，使用目标语言的“文本-音频”再次训练得到目标语言的语音合成系统。迁移学习既可以使 Tacotron2 的解码器更快地学习语音特征信息，也可以加快解码器学习目标语言与源语言较为类似的发音单元与文本单元的对应关系，还能够复用 Tacotron2 的编码器信息，加快语言模型的建立和文本信息抽取。

然而，泰语的语法结构和发音规则与汉语和英语都有明显的差别，总结泰语与汉语和英语相似的发音单元也就较为困难。因此，我们提出了一种预训练方法，只让解码器学习语音底层特征而忽略掉源语言音频和文本单元的对齐信息，并防止源语言的文本单元和对齐信息造成干扰。具体做法是训练得到一个英语语音合成系统后舍弃掉编码器的网络权重并提供泰语“文本-音频”训练模型，这样能够提供语音共同的特征信息，使解码器只需要专注于学习编码器提供的泰语的文本单元和音频的对应关系，同样在第 3 节中设计的实验证实了所提方法的有效性。

2.4 泰语文本嵌入方法

泰语是一种拼音文字，它的基本组成单元是字符。泰语的书写顺序为自左而右书写，一般不使用标点符号。泰语中的字符可类比成英语中的 26 个字母。根据泰文工业标准 620-2533[14]，泰语字符的 Unicode 编码范围为 0E00-0E7F 代码点，共 128 个代码位，目前已使用 87 个代码位，其他 41 个代码位作为保留代码位暂未使用，即可以把泰语看成由 87 个字符组合而成。字符又可以组成泰语的 44 个辅音（现代泰语实际用到的辅音有 42 个）、32 个元音（现代泰语实际用到的元音有 30 个）、5 个声调，以及一些特殊符号。泰语中辅音又分为中辅音、高辅音和低辅音，元音又分为单元音、复合元音和特殊元音。元音字母可以在辅音字母的前后出现，还可以出现在辅音字母的上下部位，甚至可以把辅音字母夹在中间。泰语中共有 4 个声调符号，标在辅音的右上方，第一声调不标符号。泰语中的元音和辅音相拼时，可以根据声调发出不同的音，但是，发音必须看辅音。除元音、辅音、声调外，泰语中还存在一些不发音的特殊符号。需要强调的是，泰语中字符的排列顺序可能会和实际发音的顺序不相匹配。

在英语 Tacotron2 系统中，一般会直接采用英语中的字母和标点符号等作为文本嵌入单元。在汉语 Tacotron2 系统中，除标点符号外，可以将文字转换为汉语拼音后直接使用拉丁字母嵌入，也可以将拉丁字母进一步组合成声母、韵母作为基本嵌入单元[15]，在汉语语音合成的研究论文中，往往只会采用其中一种文本嵌入方式，不会对比两种方法的优劣。汉语和英语中将字母和标点符号等作为最小嵌入单元往往最容易实现，且每个符号都是由单个字符组成。由于标准泰语可以由 87 个字符完全表示，因此我们提取了泰语中的 87 个字符作为基本嵌入单元实现了泰语字符嵌入，文本处理流程示例如图 2 所示。采用这种嵌入方法的系统被称为字符嵌入系统。

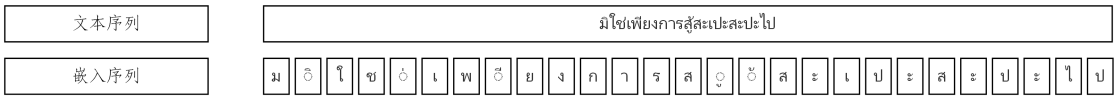


图 2 字符嵌入图例

Fig. 2 Example of character embedding

我们认为可以将泰语字符组合成元音、辅音、声调和一些特殊符号嵌入系统，与汉语中的声韵母嵌入类似。但是，在泰语中确认元辅音单元比汉语中确认声韵母单元更为困难，因为泰语中的元辅音单元并不是按照字符在文本中的排列顺序进行组合的。实现这一嵌入方式的流程示例如图 3 所示，首先对泰语进行分词，参考文献[16]中对泰语罗马化编码方案将泰语罗马化并确定了 60 个不同的嵌入单元，在 Tacotron2 的文本处理环节中使用最大正向匹配算法去匹配泰语中的元音、辅音、声调和特殊符号。采用这种嵌入方法的系统被称为元辅音嵌入系统。

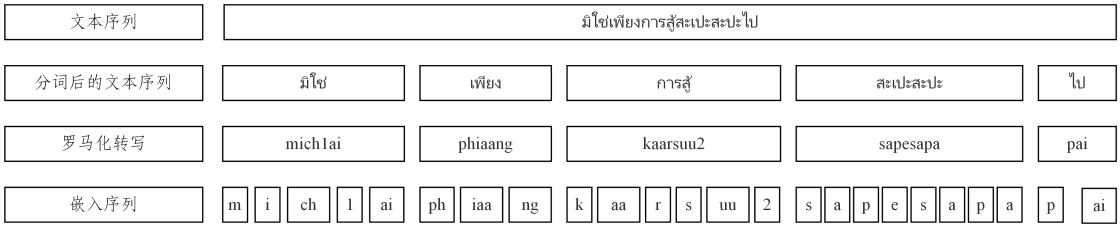


图 3 元辅音文本嵌入图例

Fig. 3 Example of vowel consonant embedding

采用字符嵌入系统的优势是对文本加入的人为约束更少，是语言最原始的特征表示，劣势是字符的排列顺序和实际发音的顺序并不完全一致，可能会对系统造成干扰。采用元辅音嵌入系统的优势是使文本发音顺序和文本嵌入单元得到对应，并且有效嵌入单元由 87 个减少到 60 个，可以通过注意力机制更好地学习到文本音频的对应关系，缺点是对前端文本施加的人为约束较多，也可能会对系统造成干扰。因此，本文在第 3 节设计了实验来对比两种嵌入方式。

3 实验

3.1 实验数据与平台

本文实验中使用的泰语数据集时长 8.08 h，由一位母语为泰语的女性说话人录制，录音人口音为纯正的泰语，文本音频对共 4 983 句，其中训练集 4 823 句，验证集 110 句，测试集

50 句。音频采样率为 22050 Hz、16 位 PCM 编码，前后含有 50 ms 的静音段。实验中使用的英语数据集来自于 LJspeech[17]，同样为单一女性说话人录制，时长为 23.9 h，我们从中选取了 12 h 的语料，并把音频采样率统一为 22050 Hz。

所有实验的代码均采用 Python 编写，基于 PyTorch 深度学习框架实现，使用一块英伟达 3090 显卡来训练，所有实验的模型的训练批次都设置为 32，共训练 500 个 epoch。实验中使用了 WaveGlow[18] 声码器将声学参数转换为语音波形。

3.2 实验设计

本文提出的交替训练和预训练是针对 Tacotron2 系统的通用方法，它们和文本嵌入方式以及语言本身特性并无关系，只需要通过实验来验证这两种方法的有效性。设计的 6 组实验主要是为了对比和研究在泰语语音合成系统中，字符嵌入和元辅音嵌入系统的性能差距，实验名称与方法如表 1 所列。

表 1 实验名称与方法

Table 1 Experiment names and methods

实验名称	实验方法
Experiment 1	字符嵌入＋Tacotron2 基线系统
Experiment 2	元辅音嵌入＋Taocotron2 基线系统
Experiment 3	字符嵌入＋交替训练
Experiment 4	元辅音嵌入＋交替训练
Experiment 5	字符嵌入＋预训练＋交替训练
Experiment 6	元辅音嵌入＋预训练＋交替训练

3.3 评测方法

对于每个实验，测试集都为之前随机挑选的 50 句泰语文本和真实音频，借助合成语音的注意力对齐图和平均意见得分(Mean Opinion Score，MOS)作为评测的标准。在 Tacotron2 系统中，注意力对齐图是体现合成质量的一个指标，它表示输入的文本序列和合成的语音帧之间的对应关系。注意力对齐图横轴代表语音帧，纵轴代表音素(字符)，对齐线最理想的状态是呈对角线趋势，代表解码器生成的语音帧与输入文本一一对应。

MOS 评分即依靠人的听觉印象来对听到的语音进行评价打分。我们邀请到 10 位泰语专业的在读硕士研究生对测试集的合成语音进行打分，最后取平均分进行统计，评分细则如表 2 所列。

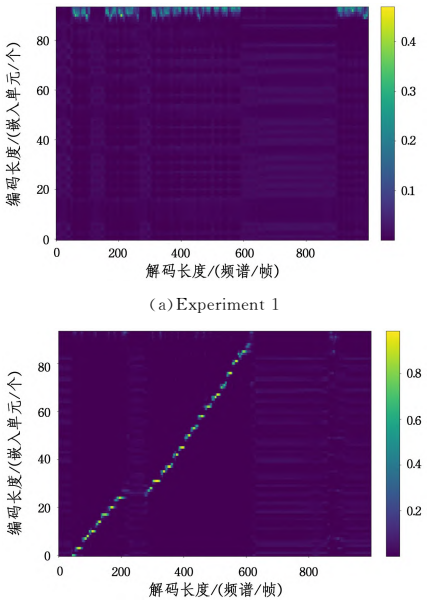
表 2 MOS 评分标准

Table 2 MOS criteria

等级	MOS 值	语音评价
优	5.0	十分自然
良	4.0	比较自然
中	3.0	可以接受
差	2.0	不自然
劣	1.0	不能接受

3.4 结果与分析

首先在测试集中随机挑选一句文本通过 Experiment 1 和 Experiment 3 合成语音的注意力对齐图(见图 4)来验证所提出的交替训练方法的有效性。可以看出，受 Tacotron2 暴露偏差的影响，Experiment 1 的模型无法合成与文本对应的语音，而 Experiment 2 采用交替训练则缓解了这一问题，合成音频开始与文本对应，并且有向对角线靠近的趋势。本次实验设置的式(1)中的 n 为 500，t_1 为 100，t_2 为 250，即初始时 $P(t)$ 为 0.2，在第 100 个 epoch 时 $P(t)$ 逐渐增加，在第 250 个 epoch 增加到 0.5 后，直至第 500 个 epoch 保持 0.5 不变得到的效果最为明显。



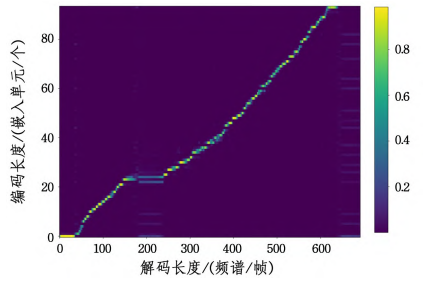
(a) Experiment 1

(b) Experiment 3

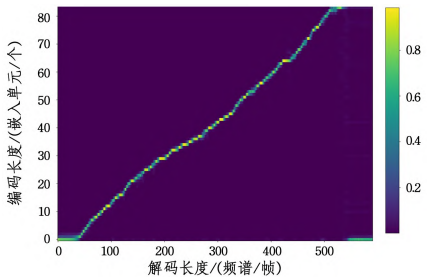
图 4 实验 1 和实验 3 的对齐图

Fig. 4 Alignment figures of experiments 1 and 3

接下来仍然选取相同的文本，通过 Experiment 5 的对齐图(见图 5)来验证所提出的预训练方法的有效性。与 Experiment 2 的对齐图相比，Experiment 5 的对齐线不仅断点的数量大大减少，而且合成语音末尾的静默段也减少了，更接近原始音频的真实时长。



(a) Experiment 5



(b) Experiment 6

图 5 实验 5 和实验 6 的对齐图

Fig. 5 Alignment figures of experiments 5 and 6

由此可以得出结论，本文提出的交替训练方法和预训练方法可以在低资源条件下提升语音合成系统的性能。

Experiment 6 也采用了和之前的 3 组实验同样的合成文本，它的注意力对齐图(见图 5)最接近对角线且断点数量最少，与 Experiment 5 采用字符嵌入的对齐图相比有明显优势。因此，我们得出采用元辅音嵌入的系统的性能要优于采用字符嵌入的系统的性能这一初步结论。

为了验证我们的初步结论，本文对 Experiment 1－Experiment 6 的模型在测试集上的合成语音进行了 MOS 评测，结果如表 3 所列。

表 3　MOS 评测结果

Table 3　MOS evaluation results

音频来源	MOS 值
原始音频	4.68
Experiment 1	1.71
Experiment 2	1.83
Experiment 3	2.66
Experiment 4	2.83
Experiment 5	3.38
Experiment 6	3.95

由表 3 可以得出，相同条件下元辅音嵌入系统的平均 MOS 得分均高于字符嵌入系统，在采用了本文提出的交替训练方法和预训练方法后，元辅音嵌入系统的平均 MOS 得分最高。可以得出结论：针对基于 Tacotron2 的泰语语音合成系统，同等条件下元辅音嵌入的效果优于字符嵌入的效果。我们认为，这可能是因为采用元辅音嵌入减少了文本嵌入单元的数量，并能够为 Tacotron2 编码器提供更多的语法信息。

结束语　本文以 Tacotron2 为基线系统构建泰语语音合成系统，提出了一种交替训练的方法来解决 Tacotron2 的暴露偏差问题，并提出了一种预训练的方法来提升系统的性能，设计并探讨了字符嵌入和元辅音嵌入的差异，通过对比实验得出在同等条件下采用元辅音嵌入的系统优于字符嵌入系统的结论。接下来，我们将探索如何对 Tacotron2 的编码器进行改进，以进一步提升系统的性能。

参 考 文 献

[1] ARIK S Ö, CHRZANOWSKI M, COATES A, et al. Deep Voice: Real-Time Neural Text-to-Speech[C]// International Conference on Machine Learning. Singapore: PMLR, 2017: 195-204.

[2] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards End-to-End Speech Synthesis[C]// Proceedings of Conference of the International Speech Communication Association. Stockholm, 2017: 4006-4010.

[3] SHEN J, PANG R, WEISS R J, et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 4779-4783.

[4] REN Y, HU C, TAN X, et al. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech[J]. arXiv: 2006. 04558, 2020.

[5] PING W, PENG K, GIBIANSKY A, et al. Deep Voice 3: 2000-Speaker Neural Text-to-Speech[C]// Proceedings of the 3rd International Conference on Learning Representations. Toulon, 2017: 1-15.

[6] WUTIWIWATCHAIC, HANSAKUNBUNTHEUNGC, RUGCHATJAROEN A, et al. Thai text-to-speech synthesis: a review [J]. Journal of Intelligent Informatics and Smart Technology, 2017, 2(2): 1-8.

[7] CHOMPHAN S, KOBAYASHI T. Implementation and Evaluation of An HMM-Based Thai Speech Synthesis System[C]// Proceedings of Conference of the International Speech Communication Association. Antwerp, 2007: 2849-2852.

[8] TESPRASIT V, CHAROENPORNSAWAT P, SORNLERTLAMVANICH V. A context-sensitive homograph disambiguation in Thai text-to-speech synthesis[C]// Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers. Association for Computational Linguistics. Edmonton, 2003: 103-105.

[9] WAN V, LATORRE J, CHIN K K, et al. Combining multiple high quality corpora for improving HMM-TTS[C]// Proceedings of Conference of the International Speech Communication Association. Portland, 2012: 1135-1138.

[10] LIU R, SISMAN B, LI J, et al. Teacher-student training for robust Tacotron-based TTS[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 6274-6278.

[11] LIU R, YANG J, LIU M. A New End-to-End Long-Time Speech Synthesis System Based on Tacotron2[C]// Proceedings of the 2019 International Symposium on Signal Processing Systems. Beijing, 2019: 46-50.

[12] FAHMY F K, KHALIL M I, ABBAS H M. A Transfer Learning End-to-End Arabic Text-to-Speech (TTS) Deep Architecture[C]// Workshop on Artificial Neural Networks in Pattern Recognition. Winterthur: Springer, 2020: 266-277.

[13] XU J, TAN X, REN Y, et al. LRSPEECH: Extremely Low-Resource Speech Synthesis and Recognition[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. USA, 2020: 2802-2812.

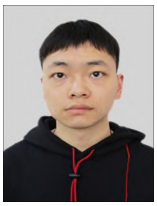
[14] TIS 620-2533, Standard for Thai Character Codes for Computers [S]. Bangkok: Thai Industrial Standard Institute, 1990.

[15] LIU J, XIE Z, ZHANG C, et al. A Novel Method for Mandarin Speech Synthesis by Inserting Prosodic Structure Prediction into Tacotron2 [J]. International Journal of Machine Learning and Cybernetics, 2021, 12(10): 2809-2823.

[16] LIU H J, YANG J, XIONG Y J, et al. Implementation of Word Segmentation and Romanization for Thai Text[C]// NCMMSC'2013. Guiyang, 2013.

[17] KEITH I, LINDA J. The LJ Speech Dataset[OL]. https://keithito.com/LJ-Speech-Dataset/.

[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. Brighton: IEEE, 2019: 3617-3621.



CAI Haoran, born in 1997, postgraduate. His main research interests include speech synthesis, recognition and understanding.

YANG Jian, born in 1964, Ph.D, professor. His main research interests include speech synthesis, recognition and understanding.