南京工業大學

毕业设计(论文)外文文献翻译

学生姓名： 居泽灵 **学 号：** 201921008072

所在学院： 计算机科学与技术学院

专　　业： 计算机科学与技术

设计(论文)题目： 基于 2D 地图探索的轻量级游戏渲染的设计与实现

指导教师： 杜振龙

2022 年 2 月 27 日

PaddleSpeech:一个易于使用的一体化语音工具包

Hui Zhang1 , Tian Yuan1 , Junkun Chen3 , Xintong Li2 , Renjie Zheng2 , Yuxin Huang1 , Xiaojie Chen1 , Enlei Gong1 , Zeyu Chen1 , Xiaoguang Hu1 , Dianhai Yu1 , Yanjun Ma1 , Liang Huang2,3

1Baidu Inc., Beijing, China

2Baidu Research, Sunnyvale, CA, USA

3Oregon State University, Corvallis, OR, USA

{zhanghui41, yuantian01, xintongli, renjiezheng, huangyuxin, chenxiaojie06, gongenlei, chenzeyu01}baidu.com

摘要

PadderSpeech 是一个开源的多功能语音工具包。它旨在通过提供易于使用的命令行界面和简单的代码结构，促进语音处理技术的开发和研究。本文描述了 PadleSpeech 的设计理念和核心架构，以支持几个基本的语音到文本和文本到语音任务。PadleSpeech 在各种语音数据集上实现了极具竞争力或最先进的性能，并实现了最流行的方法。它还提供了配方和预训练模型，以快速再现本文中的实验结果。

1 介绍

语音处理技术使人类能够 直接与计算机通信，这是 这是巨大应用程序的重要组成部分 作为智能家居设备(Hoy,2018)，自主以及同声翻译(Zheng et al., 2020)。开源工具包降低了语音处理技术的应用和研究障碍，从而促进了语音处理技术的发展领域(Young et al., 2002; Lee et al., 2001; HugginsDaines et al., 2006; Rybach et al., 2011; Povey et al., 2011; Watanabe et al., 2018; Han et al., 2019; Wang et al., 2020; Ravanelli et al., 2021; Zhao et al., 2021)。

然而，目前流行的语音处理工具包会认定他们的用户是有经验的从业者或研究人员，因此初学者在开发他们令人兴奋的应用程序时可能会感到困惑。例如，使用 Kaldi(Povey et al., 2011)构建新的语音应用程序原型，用户必须能够轻松地阅读和修改使用 Bash、Perl 和 Python 脚本编写的决策，并精通 C++以洞悉其实现。最近的工具包例如 Fairseq S2T (Wang et al., 2020)和 NeurST(Zhao et al., 2021)，在建立在通用深度学习库上变得更加灵活。但它们复杂的代码风格也导致学习费时，而且很难从一个迁移到另一个。因此，我们开发了 PaddleSpeech，提供命令行界面和便携式功能，使每个人都能开发与语音相关的应用程序。

值得注意的是，中文社区有许多开发者渴望为社区做出贡献。然而几乎所有的深度学习库，例如 Pytorch(Paszke et al., 2019) 和 Tensorflow(Abadi et al., 2016)，主要都是针对英文社区，大大增加了中文开发者的难度。PaddlePaddle 作为唯一面向中英文社区的全功能开源深度学习平台，累计超过

500k 提交，476k 模型，被 157k 企业使用。因此，我们期待与 PaddlePaddle 一起开发的 PaddleSpeech 能够消除英语和中文社区之间的障碍，从而促进语音技术和应用的发展。

为行业开发语音应用程序与在学术界进行研究不是同一种情况。研究论文主要集中在开发新的模型，以更好地执行特定的数据集。但是，在应用语音产品时通常不存在干净的数据集。因此，PaddleLanguage 为原始音频提供了即时预处理，使 PaddleLanguage 可以直接用于面向产品的应用程序。值得注意的是一些预处理方法在 PaddleLanguage 中是独占的，例如基于规则的中文文本到语音前端，能够显著提高合成语音的性能。

性能是所有应用程序的基石。PaddleSpeech 达到了在各种常用基准上最先进的亦或是具有竞争力的表现，如表 1 所示。

Task	Description	Techniques	Datasets
Sound Classification	*Label sound class*	Finetuned PANN (Kong et al., 2020b)	ESC-50 dataset (Piczak, 2015)
Speech Recognition	*Transcribe speech to text*	Deepspeech2 (Amodei et al., 2016) Conformer (Zhang et al., 2020) Transformer (Zhang et al., 2020)	Librispeech (Panayotov et al., 2015) AISHELL-1 (Bu et al., 2017)
Punctuation Restoration	*Post-add punctuation to transcribed text*	Finetuned ERNIE (Sun et al., 2019)	IWSLT2012-zh (Federico et al., 2012)
Speech Translation	*Translate speech to text*	Transformer (Vaswani et al., 2017)	MuST-C (Di Gangi et al., 2019)
Text To Speech	*Synthesis speech from text*	**Acoustic Model** Tacotron 2 (Shen et al., 2018) Transformer TTS (Li et al., 2019) SpeedySpeech (Vainer and Dušek, 2020) FastPitch (Łańcucki, 2021) FastSpeech 2 (Ren et al., 2020) **Vocoder** WaveFlow(Ping et al., 2020) Parallel WaveGAN (Yamamoto et al., 2020) MelGAN (Kumar et al., 2019) Style MelGAN (Mustafa et al., 2021) Multi Band MelGAN (Yang et al., 2021) HiFi GAN (Kong et al., 2020a)	CSMS (DataBaker) AISHELL-3 (Shi et al., 2020) LJSpeech (Ito and Johnson, 2017) VCTK (Yamagishi et al., 2019)

Table 1: List of speech tasks and corpora that are currently supported by PaddleSpeech.

我们在本文中的主要贡献是两方面的

我们将介绍如何设计 PaddleSpeech 以及它支持哪些功能。

我们提供实现和可重现的实验细节，产生在各种任务中最先进的亦或有竞争力的表现。

2 Design of PaddleSpeech

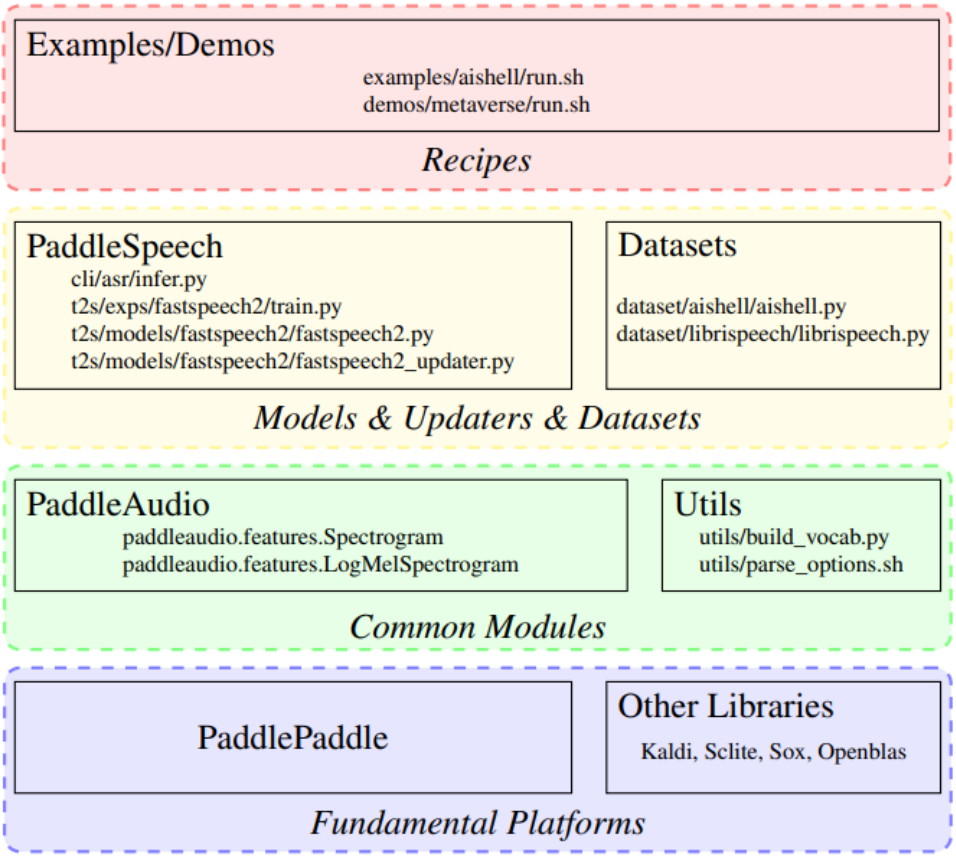

Figure 1: Software architecture of PaddleSpeech.

图 1 显示了 PaddleSpeech 的软件架构。作为一个易于使用的语音处理工具包，PaddleSpeech 提供了许多完整的决策来执行各种与语音相关的任务，并演示命令行界面的使用。熟悉顶层就足够构建与语音相关的应用程序。

第二层面向的是语音和语言处理方面的研究人员。PaddleSpeech 的设计理念以模型为中心，简化语音处理方法的学习和开发。对于特定的方法，特定模型的所有计算都包含 PaddleSpeech/<task>/models/<model>下的两个文件中。

PaddleSpeech 已经实现了大多数常用的和性能良好的模型。模型体系结构在一个独立的文件中实现，该文件由方法命名。相应的训练步骤和评估步骤在另一个更新程序文件中实现。通常，阅读或洞悉这两个文件就足以理解或设计模型。更高级的洞悉关于更精细的数据处理或更复杂的训练/评估循环，也可在 PaddleSpeech/任务></经验/><模型>。原始数据集可通过对应 dataset/<dataset>/中的脚本获取。PaddleSpeech 支持分布式多 gpu 高效率训练。

标准模块例如音频和文本特性转换和实用程序脚本在第三层中作为库实现。PaddleSpeech 的后端主要是 PaddlePaddle，部分函数来自第三方库，如第四层

所示。PaddleSpeech 提供了多种方法，使用 PaddleAudio 和 Kaldi 从原始音频中提取多种类型的语音特征，例如频谱图和滤波器组，可以根据任务的需求进行更改。

3 实验

在这部分，我们将 PaddleSpeech 中的模型与其他流行实现在五个与语音相关的任务中的性能进行比较，包括声音分类、语音识别、发音、语音翻译和语音合成。工具箱可以在大多数任务上达到 SOTA。本节中的所有实验都包括数据准备、评估指标和提高可重复性的实施细节。

3.1 声音分类

声音分类是一项识别特定声音的任务，包括语音命令(Warden,2018)，环境声音(Piczak, 2015)，识别乐器(Engel et al.,2017)，寻找鸟鸣(Stowell et al.,2018)，情感识别(Xu et al.,2019)和说话人验证(Liu et al.,2018)。

数据集

在这部分中，我们分析了在 ESC-50 数据集(Piczak, 2015)上的声音分类中 PaddleSpeech 的性能。ESC-50 数据集是 2000 个 5 秒环境音频记录的标记集合，包括 50 个声音事件，如“狗”、“猫”、“呼吸”和“烟花”，每个事件 40 个录音。

数据处理

首先，我们将所有音频录音重新采样到 32kHz，并将其转换为单声道以与 AudioSet 上训练的 PANNs 一致(Kong et al.,2020b)。然后，我们通过对汉明窗大小为 1024，跳跃步长为 320 样本的波形进行短时傅里叶变换，将录音转换为 log mel 频谱图。这种配置导致每秒 100 帧。继 Kong et al.(2019)之后，我们使用 64mel 滤波器组来计算 log mel 谱图。

实现

PANNs (Kong et al.,2020b)是用于音频相关任务的预训练 CNN 模型之一，其特征在于使用 AudioSet (Gemmeke et al.,2017)。PANNs 对于只提供有限数量的训练片段的任务很有帮助。在这种情况下，我们为环境声音分类任务微调 PANN 的所有参数。所有参数都从 PANN 初始化，除了最后的全连接层是随机初始化的。具体来说，我们分别用 6 层、10 层和 14 层实现了 CNNs(Kong et al.，2020b)。

结果

我们报告了在 ESC-50 数据集上的 5 倍交叉验证精度值。如表 2 所示，PANNs-CNN14 达到 0.9500 5 倍交叉验证精度，与目前最先进的方法相当(Gong et al.,2021)。

Model	Accuracy
AST-P (Gong et al., 2021)	95.6 ± 0.4
PANNs-CNN14	95.00
PANNs-CNN10	89.75
PANNs-CNN6	88.25

Table 2: 5-fold cross validation accuracy of ESC-50.

3.2 自动语音识别

自动语音识别(ASR)是一种将音频内容转录为相同语言的文本的任务。

数据集

我们在 Librispeech(Panayotovet al.,2015)和 Aisell -15(Bu et al.,2017)两个主要数据集上进行 ASR 实验。Librispeech 包含 1000 小时语音数据。整个数据集分为 3 个训练集(100h 清洗，360h 清洗，500h 其他)，2 个验证集(清洗，其他)和 2 个测试集(清洗，其他)。Aishell 包含 178 小时语音数据。来自中国不同口音地区的 400 名演讲者参与了录音。数据集分为训练集(340 人)、验证集(40 人)和测试集(20 人)。

数据预处理

Deepspeech 2 采用两种英语的字符级词汇和普通话任务。对于其他模型，我们使用汉字级词汇。并用 SentencePiece(Kudo And Richardson, 2018)对英语文本进行预处理。这两种数据集都增加了四个额外的字符，分别是<’>, <space>, <blank> and <eos>。对于倒谱均值和方差归一化(CMVN)，选择训练集的一个子集或全部子集，用于计算特征均值和标准误差。对于特征提取，我们实现了几种方法，如线性频谱图，滤波器组和梅尔频率倒谱系数。目前，Deepspeech 2 模型使用线性谱图或滤波器组，但 Transformer 和 Conformer 模型使用滤波器组。为了公平的比较，我们在 Transformer 中加入了额外的三维音高特征，以与 ESPnet 保持一致。

实现

我们实现了流媒体和非流媒体 Deepspeech 2 (Amodei et al., 2016)。非流模型有 2 个卷积层和 3 个 LSTM 层。流媒体 该模型有 2 个卷积层和 5 个 LSTM 层。Conformer 和 Transformer 模型是根据 Zhang et al(2020)实现的，具有 12 个编码器层和 6 个解码器层。

结果

我们分别报告了 Librispeech(英语)和 Aishell(普通话)语音识别的单词错误率(WER)和字符错误率(CER)。如表 3 所示，Conformer 和 Transformer 优于 Deepspeech 2。与相关工作相比，我们的最佳模型在两个数据集上都实现了相当的性能。

Data	Model	Streaming	Test Data	Language Model	CER	WER
Aishell	WeNet Conformer†* (Yao et al., 2021)	✓			5.45	-
	WeNet Conformer† (Yao et al., 2021)				4.61	-
	WeNet Transformer† (Yao et al., 2021)				5.30	-
	ESPnet Conformer† (Inaguma et al., 2020)				5.10	-
	ESPnet Transformer† (Inaguma et al., 2020)				6.70	-
	SpeechBrain Transformer† (Ravanelli et al., 2021)				5.58	-
	Deepspeech 2	✓		char 5-gram	6.66	-
	Deepspeech 2			char 5-gram	6.40	-
	Transformer				5.23	-
	Conformer*	✓			5.44	-
	Conformer				4.64	-
Librispeech	WeNet Conformer† (Yao et al., 2021)		test-clean		-	2.85
	SpeechBrain Transformer† (Ravanelli et al., 2021)		test-clean	TransformerLM	-	2.46
	ESPnet Transformer† (Inaguma et al., 2020)		test-clean	TransformerLM	-	2.60
	Deepspeech 2		test-clean	word 5-gram	-	7.25
	Conformer		test-clean		-	3.37
	Transformer		test-clean	TransformerLM	-	2.40

† denotes the results are reported in their public repositories.
* denotes the results are streaming with chunk size 16.

Table 3: WER/CER on Aishell, Librispeech for ASR Tasks.

3.3 标点符号恢复

标点恢复是 ASR 系统的后处理问题。提高转录文本对人类读者的可读性和促进下行 NLP 任务至关重要。

数据集

我们在包含 150k 汉语标点句子的 IWSLT2012-zh6 数据集上进行了实验。在这个任务中，我们选择逗号、句号和问号作为恢复目标，因此在训练模型之前，我们用这三个标记替换其他标点符号。我们将数据分为训练集、验证集和测试集，分别使用 147k、2k 和 1k 样本。

实现

我们将标点恢复问题定义为具有四个目标类的序列标记任务，包括 EMPTY、COMMA、PERIOD 和 QUESTION (Nagy et al., 2021b)。ERNIE (Sun et al., 2019)作为一种预训练的语言模型，在五个中文自然语言处理任务上取得了最新的结果，包括:自然语言推理、语义相似性、命名实体识别、情感分析和问题回答。因此，我们为这项任务调整了 ERNIE 模型。更具体地说，所有参数都从 ERNIE 预训练的模型初始化，除了最后的共享全连接层是随机初始化的。

结果

我们在 IWSLT2012-zh 数据集上报告 F1-score 值。如表 5 所示，我们的 ERNIELinear 模型总体 f1 得分为 0.6331，与之前的工作(Nagy et al., 2021a)相当。

model	COMMA	PERIOD	QUESTION	Overall
BERTLinear†	0.4646	0.4227	0.7400	0.5424
BERTBiLSTM†	0.5190	0.5707	0.8095	0.6330
ERNIELinear	0.5142	0.5447	0.8406	0.6331

† denotes the results come from our reproduced models.

Table 5: F1-score values on IWSLT2012-zh dataset.

3.4 语音翻译

语音翻译，即将源语言的语音翻译成另一种语言的文本，有利于人类交流。

数据集

在这部分，我们分析了使用 PaddleSpeech 在 MuST-C 数据集(Di Gangi et al., 2019)上使用 8 种不同的语言翻译对进行语音到文本翻译的性能，这些语言翻译对采用英语语音作为源输入。

实现

我们使用 Kaldi (Povey et al., 2011)处理原始音频，并使用 25ms 的窗口大小和 10ms 的步长提取 80 维的 log-mel 滤波器组。文本首先使用 Moses tokenizer7 进行标记，然后由 SentencePiece (Kudo and Richardson, 2018)使用每个语言对大小为 8K 的联合词汇表进行处理。我们使用 Transformer (Vaswani et al., 2017)作为语音翻译实验的基础架构。具体来说，Transformer 模型有 12 层编码器层和 6 层解码器层，编码器层遵循 2 层二维卷积，核大小为 3，步幅大小为 2。每一层包含 4 个注意头，大小为 256。编码器由预训练的 ASR 模型初始化。

结果

我们报告解标记区分大小写的 BLEU。如表 4 所示，与其他框架相比，PaddleSpeech 可以获得具有竞争力的结果。

Frameworks	De	Es	Fr	It	Nl	Pt	Ro	Ru
ESPnet-ST (Inaguma et al., 2020)	22.9	**28.0**	32.8	**23.8**	**27.4**	28.0	21.9	**15.8**
fairseq-ST (Wang et al., 2020)	22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3
NeurST (Zhao et al., 2021)	22.8	27.4	**33.3**	22.9	27.2	28.7	**22.2**	15.1
PaddleSpeech	**23.0**	27.4	32.9	22.9	26.7	**28.8**	**22.2**	15.4

Table 4: Case-sensitive detokenized BLEU scores on MuST-C *tst-COMMON*.

3.5 文本转语音

文本到语音(TTS)系统将给定的语言文本转换为语音。PaddleSpeech 的 TTS 管道包括三个步骤。我们首先通过文本前端模块将原始文本转换为字符/音素。然后，通过声学模型，我们将字符或音素转换为声学特征，如 mel 谱图，最后，我们通过声码器从声学特征生成波形。在 PaddleSpeech 中，文本前端是一个受专家知识启发的基于规则的模型。声学模型和声码器是可训练的。

数据集

在 PaddleSpeech 中，我们主要关注普通话和英语语音合成。我们有 CSMSC, AISHELL-3, LJSpeech, VCTK 的基准测试。由于篇幅所限，我们只在 CSMSC 上列出了实验结果，其中包含了 12 小时的演讲音频，对应 10k 个句子。

文本前端

文本前端模块用于从给定文本中提取语言特征、字符和音素。主要包括:文本分割、文本归一化(TN)、单词分割(WS)、词性标注、韵律预测和字母到音素(G2P)(见表 6)。

	Module	Result												
PaddleSpeech	*Text*	*jīn tiān* 今天 today	shì 是 is	2020/10/29	,	zuì dī 最低 lowest	wēn dù 温度 temperature	shì 是 is	-3°C	。				
	TN	今天	是	*èr líng èr líng nián shí yuè èr shí jǐu rì* 二 零 二 零 年 十 月 二十九 日 2 0 0 2 year 10 month 29 day	,	最低	温度	是	*líng xià sān dù* 零下 三 度 *negative three degree*	。				
	WS	今天 /	是 /	二零二零年 / 十月 / 二十九日	,	/ 最低	温度 /	是 /	零下 / 三度	。				
	G2P	jin1 tian1 shi4	er4 ling2 er4 ling2 nian2 shi4 yue4 er4 shi2 jiu3 ri4			zui4 di1	wen1 du4	shi4	ling2 xia4 san1 du4					
ESPnet	jin1 tian1 shi4	2020/10/29			zui4 di1	wen1 du4	shi4	-3°C						

Table 6: An example of the text preprocessing pipeline for Mandarin TTS of PaddleSpeech and ESPnet. **TN** stands for the text normalization module, **WS** stands for the word segmentation module, **G2P** stands for the grapheme-to-phoneme module. The text normalization module for mandarin of ESPnet is not able to correctly handle dates (2020/10/29) and temperatures (-3°C).

对于普通话，我们的 G2P 系统包括一个复音模块，它使用 pypinyin 和 g2pM，以及一个变调模块，它使用基于中文分词的规则。据我们所知，与其他公开发布的作品相比，我们的中文文本前端系统是最完整的。

数据预处理

PaddleSpeech TTS 使用以下模块进行数据预处理:首先，我们使用蒙特利尔强制对齐器来获取对应音素的持续时间。其次，提取 mel 谱图作为特征(Fastspeech 2 的额外音调和能量特征)。最后，对每个特征进行统计归一化。

声学模型

声学模型主要分为自回归模型和非自回归模型。自回归模型的解码依赖于每一步之前的预测，这导致更长的推理时间，但相对更好的质量。而非自回归模型是并行生成输出，因此推理速度较快，但生成的结果质量相对较差。

如表 1 所示，PaddleSpeech 实现了以下常用的自回归声学模型:Tacotron 2 和 Transformer TTS，以及非自回归声学模型:SpeedySpeech, FastPitch 和

FastSpeech 2。

语音编码器

如表 1 所示，PaddleSpeech 实现了以下声码器:WaveFlow, Parallel WaveGAN, MelGAN, Style MelGAN, Multi - Band MelGAN 和 HiFi GAN。

实现

FastSpeech 2 的 PadddleSpeech TTS 实现在 FastPitch 的基础上做了一些改进，使用 MFA 来获得强制对齐(FastSpeech 原论文使用的是 Tacotron 2)。值得注意的是，声学模型的语音特征参数和一个 TTS 管道的声码器应该是相同的。详细的设置可以在 CSMSC 数据集上的样例配置文件中找到。

结果

我们在表 7 中报告了自然度评估的平均意见评分(MOS)。我们使用 crowdMOS 工具包(Ribeiro et al., 2011)，其中 14 个普通话样本(见附录 A)来自这 7 个不同的模型，展示给 Mechanical Turk 上的 14 名工作者。如表 7 所示，PaddleSpeech 在普通话 TTS 上的表现明显优于 ESPnet。主要原因是 PaddleSpeech TTS 具有更好的文本前端，如表 6 所示。与其他模型相比，Fastspeech 2 结合 HiFi GAN 可以达到最好的效果。

	Acoustic Model	Vocoder	MOS ↑
ESPnet	Fastspeech 2	PWGAN	2.55 ± 0.19
PaddleSpeech	Tacotron 2	PWGAN	3.69 ± 0.11
	Speedyspeech	PWGAN	3.79 ± 0.09
	Fastspeech 2	PWGAN	4.25 ± 0.09
	Fastspeech 2	Style MelGAN	4.32 ± 0.10
	Fastspeech 2	MB MelGAN	4.43 ± 0.09
	Fastspeech 2	HiFi GAN	4.72 ± 0.08

4 总结

本文介绍了 PaddleSpeech，一个开源的，易于使用的，全合一的语音处理工具包。我们阐述了该工具包背后的主要设计理念，以进行各种与语音相关的可访问任务的开发和研究。大量可重复的实验和比较表明，在标准基准测试中，PaddleSpeech 与最流行的模型相比，达到了最先进的或具有竞争力的性能。

5 致谢

我们衷心感谢匿名审稿人提出的宝贵意见和建议。国家重点研发计划项目(2020AAA0103503)资助。