Computer Science and Application 计算机科学与应用, 2021, 11(4), 885-891
Published Online April 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.114091

Hans汉斯

基于迁移学习的端到端发音检错研究

高文明，吴怡之，魏新享

东华大学，上海
Email: gaowm219@163.com, yz_wu@dhu.edu.cn, wxx.email@qq.com

收稿日期：2021年3月15日；录用日期：2021年4月10日；发布日期：2021年4月21日

摘　要

自动发音检错是为了满足第二语言学习者发音练习的需求，而先进的自动发音检错系统通常取决于声学模型识别率。随着深度学习技术的发展，端到端声学模型算法已经逐渐成熟，为发音检错算法研究提供的新思路。本文首先构建了基于连接时序分类(Connectionist Temporal Classification, CTC)算法的端到端发音检错声学模型架构。其次，基于二语迁移现象，L2发音往往带有其母语的音素特征，本文利用迁移学习算法提高基于母语的声学模型性能，从而提高发音检错准确率。通过迁移中文母语音素特征的声学模型相比于只使用英文母语的声学模型在错误音素率上有所下降，并且训练时间减少了7.3%。在发音检错性能上检错正确率提升了2.06%。

关键词

音素，迁移学习，自动发音检错；CTC，长短记忆神经网络

Research on End-to-End Pronunciation Error Detection Based on Transfer Learning

Wenming Gao, Yizhi Wu, Xinxiang Wei

Donghua University, Shanghai
Email: gaowm219@163.com, yz_wu@dhu.edu.cn, wxx.email@qq.com

Received: Mar. 15th, 2021; accepted: Apr. 10th, 2021; published: Apr. 21st, 2021

Abstract

Automatic pronunciation error detection is to meet the needs of second language learners' pronunciation practice, and advanced automatic pronunciation error detection systems usually depend on the recognition rate of the acoustic model. With the development of deep learning tech-

文章引用: 高文明, 吴怡之, 魏新享. 基于迁移学习的端到端发音检错研究[J]. 计算机科学与应用, 2021, 11(4): 885-891. DOI: 10.12677/csa.2021.114091

nology, end-to-end acoustic model algorithms have gradually matured, providing new ideas for the research of pronunciation error detection algorithms. This paper first builds an end-to-end pronunciation error detection acoustic model architecture based on the Connectionist Temporal Classification (CTC) algorithm. Secondly, based on the phenomenon of second language transfer, L2 pronunciation often has the phoneme characteristics of its native language. This paper uses transfer learning algorithms to improve the performance of the acoustic model based on the native language, thereby improving the accuracy of pronunciation error detection. Compared with the acoustic model that only uses the native English language, the acoustic model that transfers the Chinese phoneme features has a lower error phoneme rate, and the training time is reduced by 7.3%. The correct rate of error detection in pronunciation error detection performance has increased by 2.06%.

Keywords

Phoneme, Transfer Learning, Automatic Pronunciation Error Detection, CTC, Long and Short memory Neural Network

1. 介绍

随着经济全球化的发展，越来越多的人渴望学习除母语外的第二语言。计算机辅助语言学习(Computer-Assisted Language Learning, CALL)技术[1]可实现随时随地可访问性的语言自主学习[2]。如今，市场上的 CALL 软件如雨后春般出现，可满足第二语言教学老师的匮乏，也可满足日益增长的学习者个性化的需求[3]。专注于为学习者提供 L2 发音练习的 CALL 系统通常称为计算机辅助发音训练系统(Computer-Assisted Pronunciation Training, CAPT)，CAPT 系统可以有效地处理和分析学习者说出的语音信号。然后对语音进行定量或定性评估。然后反馈给学习者其发音的评分。自动发音检错[4] (Automatic Pronunciation Error Detection, APED)模块是 CAPT 系统的核心。能够准确检测出学习者的错误发音。

最近，端到端声学模型在自动语言识别(Automatic Speech Recognition, ASR)上取得了优良成绩。端到端声学模型与传统基于隐马尔可夫模型(Hidden Markov Model, HMM) [5] [6] [7]不同，不需要在帧级别上对齐标签。另外，考虑到 APED 系统往往满足的是学习除母语外第二语言的人群，如欧洲人学习中文，日本人学习英文。基于二语迁移现象，L2 发音往往带有其母语的音素特征。本文利用迁移学习算法提高母语的端到端声学模型性能，从而提高发音检错准确率。最终本文构建基于迁移学习的端到端 APED 系统。在本文中，我们的端到端 APED 系统在第 2 节中介绍。在第 3 节中，进行实验与分析。最后，结论是在第 4 节中描述。

2. 端到端 APED

本节主要对端到端 APED 系统进行介绍。端到端声学模型主要将音频信号识别为音素序列。选择性能优异的声学模型是提升 APED 系统性能关键任务。识别率高的声学模型识别的音素序列经过序列比对算法比对，能够在识别与纠正学习者口音的绝对发音错误工作中表现出色。即通过声学模型识别的句子音素序列直接判断发音中哪些音素出现漏读，插入，替换等错误。

2.1. 特征提取

特征提取的目的是将语音波形数据转换成适合声学模型训练的声学特征参数。特征参数应具有一定的区分能力，以便声学建模单元之间能够相互区分开，实现准确的识别。音素识别中常用的特征有 Mel 频率倒谱系数(MelFrequency Cepstrum Coefficients, MFCC)和感知线性预测系数(Perceptual Linear Predictive, PLP)。MFCC 特征更符合人耳的听觉特性，在提取的过程中一般会把相位谱丢弃，使用“Mel 滤波器组”进行滤波，再使用离散余弦变换去掉维与维之间的相关性[8]。PLP 特征是一种基于听觉模型的参数，在噪声环境下性能会稍好一些。由于本文选取的语料库都是低噪的环境下录制完成的。所以选取的特征为 Mel 频率倒谱系数。MFCC 提取过程一般分为预加重，分帧，加窗，快速傅里叶变换，梅尔滤波器组，离散余弦变换。预加重，分帧，加窗是各种声学特征提取都会有的预处理过程。主要是将连续的音频模拟信号转换几十毫米一帧高频特性有所提升的数字信号。快速傅里叶变换将时域信号转成频域信号可以获得更多有效的发音信息。使用“Mel 滤波器组”进行滤波，再使用离散余弦变换去掉维度之间的相关性。

2.2. 基于 CTC 的端到端声学模型

当前，端到端自动识别模型有两种主要架构，一种是基于 CTC 算法模型，另一个是基于注意力的 Seq2seq 模型。前者更适合于时间序列建模。本文采用是基于连接时序分类(CTC)端到端模型算法。循环神经网络(Recurrent Neural Network, RNN)一般以序列数据为输入，通过网络内部的结构设计有效捕捉序列之间的关系特征，一般也是以序列形式进行输出。考虑到母语语音语料库和 L2 非母语语料库中音频数据是句子，选择一种特殊的卷积神经网络－长短记忆神经网络(Long And Short Memory Neural Network, LSTM)作为本模型神经网络[9]。相比与普通循环神经网络有效解决了长期依赖问题。LSTM 通过加入 3 个门，即遗忘门、输入门及输出门，这种独特的结构使得误差在传播过程中无需逐层归因，部分误差可以直接传递给下一层网络。输入门通过 Sigmoid 函数决定要更新的值。

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

其中 W_f 为输入门权重，b_f 为输入门偏置。f_t 的值在 0 至 1 之间。

遗忘门维护着 LSTM 网络的“状态”值。

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$\widetilde{C_t} = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{4}$$

其中 C_t 为更新上一个神经元“状态”值 C_{t-1} 所得。其中 W_i、W_c 为遗忘门权重，b_i、b_c 为遗忘门偏置。

最终输出门基于神经元状态值 C_t 要输出什么

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{6}$$

其中 W_o 为输出门权重，b_o 为输出门偏置。“状态”值经过 tanh 函数输出与输出门输出 o_t 相乘最终得到神经元输出 h_t。

CTC 的损失函数最大似然为：

$$L(S) = \sum_{(x,z)\in S} L(x,z) \tag{7}$$

$$L(x,z)=-\ln\sum_{u=1}^{|z'|}\alpha(t,u)\beta(t,u) \tag{8}$$

其中，$|z'|$ 表示 Z 对应的标签长度，$\alpha(t,u)\beta(t,u)$ 表示 t 时刻经过节点 u 的所有路径的概率和。

2.3. 基于迁移学习端到端 APED

迁移学习通过模拟人使用类比进行学习的能力，将在源领域中学到的知识迁移应用到目标领域中，放宽了对训练数据的限制，可以很好地解决标注样本量少的问题。迁移学习尤其是深度迁移学习[10]，被广泛应用于图像和文本领域的小样本学习。APED 系统往往是满足学习第二种语言的学习者需求。本文研究的是汉语母语的英语为第二语言(ChineseL2)的发音检测问题。利用迁移学习在深度神经网络上的应用，将声学模型在汉语母语，即，Chinese L1，语料库上训练后的参数迁移到训练英语母语，即，EnglishL1，语料库的过程中来。以这种方式去提升端到端 APED 性能。基于迁移学习的端到端 APED 如图 1。

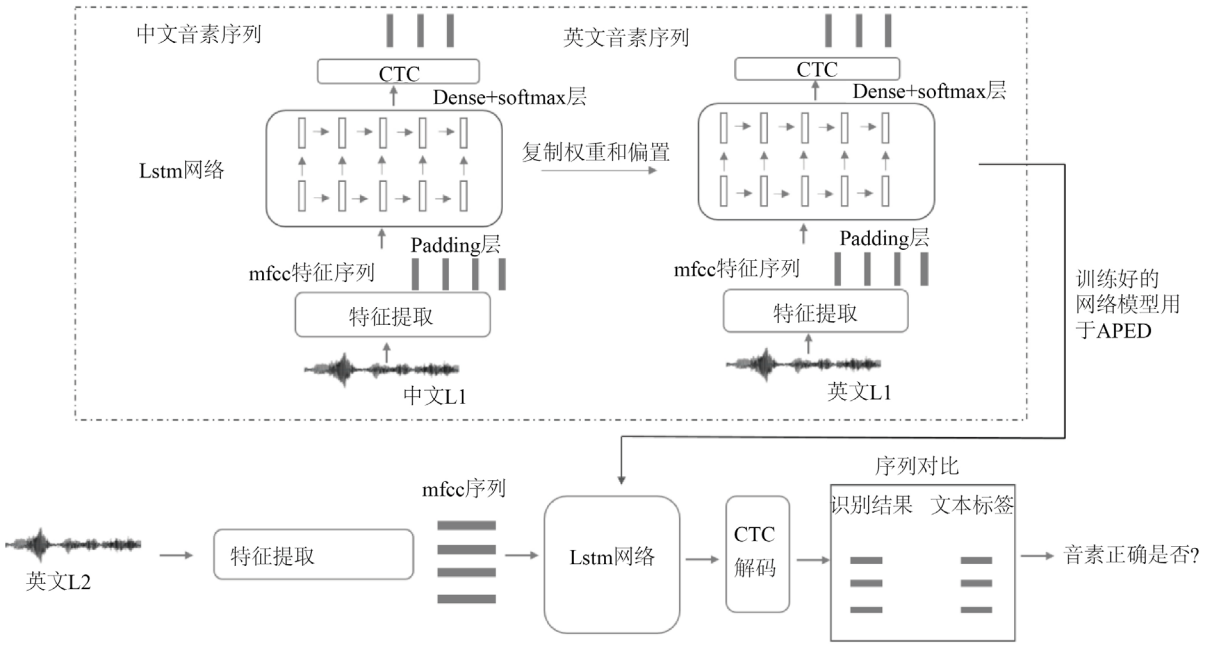

Figure 1. End-to-end APED based on transfer learning
图 1. 基于迁移学习端到端 APED

3. 实验结果与分析

3.1. 语料库

本文选择 Timit 作为 English L1 母语语音语料库[11]。是由来自美国八个主要方言地区的 630 个人每人说出给定的 10 个句子，所有的句子都在音素级别上进行了手动分割，标记。ChineseL1 母语语音语料库选取的是标贝科技有限公司录制标注的中文标准语音库，录音环境的信噪比不低于 35 dB。音频格式为 WAV，采样率为 48 KHz [12]。

另外，本文制作了小型的 English L2 语音语料库 DHU10。收集 10 名东华大学同学(4 名男性和 6 名女性)朗读的 400 句英文音频数据，并且有语言学老师对发音有错误的音素进行标注。

3.2. 实验配置

本文实验参数设置：以帧长 25 ms、帧移为 10 ms 对语言音频数据提取特征，Mel 滤波器数目为 39，

最终得到 39 维 MFCC 特征。Batch size 大小为 64，数据输入 LSTM 网络前进行 padding 操作以保证输入数据的长度一致。每层 LSTM 均包含 150 个单元，实验对选择 2 至 5 层 LSTM 进行比对。Dense 层输出维度和 softmax 层单元数为语料库的标签种类数目。声学模型的迭代次数统一为 200 次。

3.3. 基于迁移学习的模型训练

端到端声学模型训练用时为 12 小时 18 分钟，基于迁移学习端到端声学模型用时为 11 小时 24 分钟。训练时间减少了 7.3%。图二是声学模型的损失函数值随时间的下降趋势。如图 2 显示：迁移学习声学模型因为保留着中文音素的特征。所以训练前期损失函数值下降更加快。

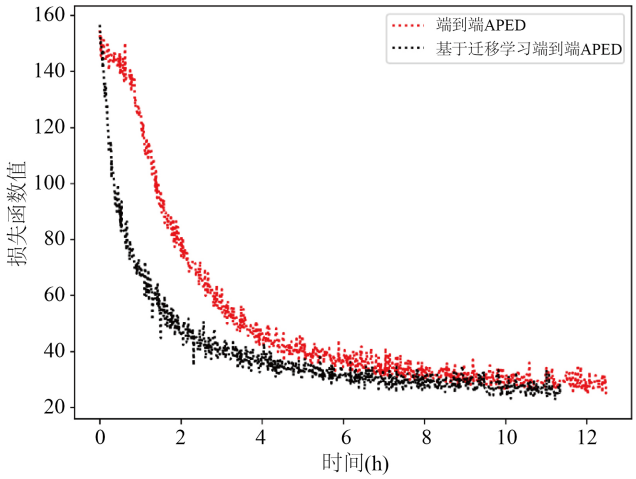

Figure 2. The loss function changes with time
图 2. 损失函数随时间变化

3.4. ASR 性能与结果

单词错误率(Word Error Rate, WER)是评估 ASR 性能最重要指标。本文主要关注音素级别的识别和检测性能。因此，我们使用音素错误率(phone error rate, PER)作为自动识别系统性能指标。

$$\mathrm{PER} = \frac{S + D + I}{N} * 100\% \tag{9}$$

其中 N 是音素总数。S，D 和 I 表示音素替换、插入错误、漏读个数，分别是通过 Needlen man-Wunsch 获得的算法比较模型识别后序列与原序列所得。

根据表 1 的结果得出。端到端声学模型在 2~4 层 LSTM 网络中，随着层数上升音素的错误率逐渐下降。在采用 5 层的 LSTM 网络时上升，是因为 English L1 语料库 TIMIT 数据不够导致的欠拟合问题。迁移学习端到端声学模型采用 5 层的 LSTM 网络时音素错误率仍然下降，一定程度上解决了样本小的问题。且在 2~5 层上的 LSTM 网络音素错误率均小于迁移学习端到端声学模型的声学模型。说明 L1、L2 母语语料库之间进行迁移的声学模型比直接使用 L2 母语语料库的更适合 APED。

3.5. APED 性能与结果

APED 的性能指标相比与 ASR 的来说较为复杂。本文绘制了图 3 的参数变量类树图。更加方便理解各个参数变量的含义。对正确的音素发音识别正确为 TA，反之为 FR。对错误的音素发音识别为正确音素为 FA，反之为 TR。

Table 1. Phoneme error rate of acoustic model
表 1. 声学模型的音素错误率

声学模型	网络层数	phone error rate(%)		
	2	3	4	5
端到端声学模型	25.48	17.62	14.30	15.22
迁移学习端到端声学模型	22.31	14.31	13.64	13.23

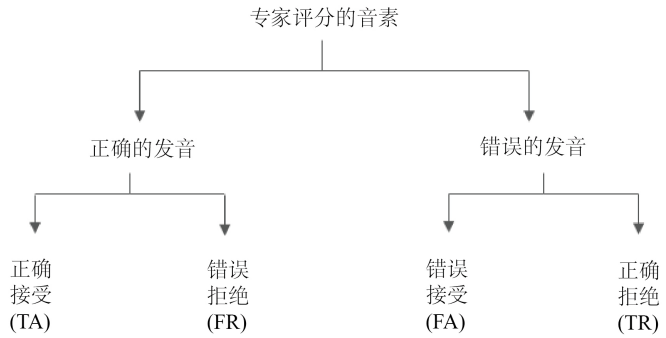

Figure 3. Parametric variable class tree diagram
图 3. 参数变量类树图

APED 的精度，召回率以及 F-measure 指标的定义分别为公式 10、11、12：

$$\text{Precision} = \frac{\text{TR}}{\text{TR} + \text{FR}} \tag{10}$$

$$\text{Recall} = \frac{\text{TR}}{\text{TR} + \text{FA}} \tag{11}$$

$$\text{F-measure} = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

系统的纠错正确率为如下公式：

$$\text{Detection Accuracy} = \frac{\text{TA} + \text{TR}}{\text{TA} + \text{FR} + \text{FA} + \text{TR}} \tag{13}$$

采用 English L2 语音语料库 DHU10 验证 APED 系统的性能。验证的性能结果为下表 2。

Table 2. APED system performance results
表 2. APED 系统性能结果

	Precision	Recall	F-measure	Detection Accuracy
端到端 APED	55.31%	79.23%	60.08%	86.17%
基于迁移学习端到端 APED	58.12%	81.05%	64.39%	88.23%

实验数据显示。基于迁移学习端到端 APED 相比端到端 APED 检错正确率提高 2.06%。经过 L1 母语语音语料迁移。APED 系统学习了学习者的母语发音特征，有利于系统对学习者进行 L2 发音检错。

4. 总结

实验表明，基于迁移学习端到端声学模型相比端到端声学模型无论在 ASR 性能上和 APED 性能上都

更优。APED 系统主要满足学习者学习第二语言的需求，结合这一特点。本文提出的基于迁移学习 APED 有着实际应用意义。

参考文献

[1] Akhtar, S., Hussain, F., Raja, F.R., *et al*. (2020) Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features. *Electronics*, **9**, 963. https://doi.org/10.3390/electronics9060963

[2] Franco, H. Neumeyer, L. Ramos, M. and Bratt, H. (1999) Automatic Detection of Phone-Level Mispronunciation for Language Learning. *Sixth European Conference on Speech Communication and Technology*, Budapest, 5-9 September 1999, 851-854.

[3] 胡文凭. 基于深层神经网络的口语发音检测与错误分析[D]: [博士学位论文]. 中国科学技术大学, 2016.

[4] Majeed, M.N., Ghazanfar, M.A., *et al*. (2019) Mispronunciation Detection Using Deep Convolutional Neural Network Features and Transfer Learning Based Model for Arabic Phonemes. *IEEE Access*, **7**, 52589-52608. https://doi.org/10.1109/ACCESS.2019.2912648

[5] Lo, W.-K., Qian, X.-J., *et al*. (2009) Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training. *Speech and Language Technology in Education* (*SLaTE* 2009), **1**, 1-4.

[6] Huang, H., Xu, H., Wang, X., *et al*. (2015) Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**, 787-797. https://doi.org/10.1109/TASLP.2015.2409733

[7] Hinton, G., Deng, L., Yu, D., *et al*. (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, **29**, 82-97. https://doi.org/10.1109/MSP.2012.2205597

[8] Davis, S. and Mermelstein, P. (1980) Comparison of Parametric Representations for Mono Syllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**, 357-366. https://doi.org/10.1109/TASSP.1980.1163420

[9] Graves, A. and Schmidhuber, J. (2005) Frame Wise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, **18**, 602-610. https://doi.org/10.1016/j.neunet.2005.06.042

[10] Oquab, M., Bottou, L., Laptev, I., *et al*. (2014) Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks. *IEEE Conference on Computer Vision & Pattern Recognition*, Columbus, 23-28 June 2014, 1717-1724. https://doi.org/10.1109/CVPR.2014.222

[11] Garofolo, J.S., Lamel, L.F., Fisher, W.M., *et al*. (1993) TIMIT Acoustic-Phonetic Continuous Speech Corpus. *Philadelphia: Linguistic Data Consortium*, LDC93S1.

[12] 标贝(北京)科技有限公司. 中文标准女声音库[EB/OL]. https://www.data-baker.com, 2016.