基于声学音素向量和孪生网络的二语者发音偏误确认*

王振宇，解焱陆*，张劲松

（北京语言大学 语言资源高精尖创新中心，北京市，100083；

*通讯作者，E-mail: xieyanlu@blcu.edu.cn）

摘要：随着自动大规模语音识别的不断发展，以自动语音识别为基础的计算机辅助发音教学也随之进步，作为传统教学方法的补充，它极大地弥补了传统教育资源不足以及传统教育方法无法及时给学习者反馈的缺陷。二语学习者的发音偏误确认和评价在计算机辅助发音训练中是较为重要的研究课题之一。针对二语者发音偏误的确认任务中缺少二语偏误发音标注问题，本文提出了一种基于声学音素向量和孪生网络的方法，将带有配对信息的成对的语音特征作为系统输入，通过神经网络将语音特征映射到高层表示，期望将不同的音素区分开。训练过程引入了孪生网络，依照输出的两个音素向量是否来自于同一类音素来调整和优化输出向量之间的距离，并通过相应的损失函数实现优化过程。结果表明使用基于余弦最大间隔距离损失函数的孪生网络获得了 89.93%的准确率，优于实验中其它方法。此方法应用在发音偏误确认任务时，不使用标注的二语发音偏误数据训练的情况下，也获得了 89.19%的诊断正确率。

关键词：发音偏误确认；音素向量； 孪生网络

中图分类号：TP391 **文献标识码：**A

Non-native Mispronunciation Verification Using Acoustic Tonal Phone Embedding and Siamese networks

Zhenyu Wang, Yanlu Xie, Jinsong Zhang*

(Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing, China,100083; * Corresponding author, E-mail: xieyanlu@blcu.edu.cn)

Abstract: With the continuous development of automatic speech recognition, the computer-aided pronunciation teaching based on automatic speech recognition has also progresses. As a supplement to traditional teaching methods, it greatly makes up for deficiency of traditional educational resources and inability of traditional educational methods to give feedback to learners in time. Pronunciation errors verification and evaluation of second language (L2) learners is one of the most important research topics in computer assisted pronunciation training. In view of the mispronunciation verification task being short of labeled mispronunciation speech data, a method based on acoustic phone embedding and Siamese network is proposed in this paper. A pair of acoustic phone segments with a pair-wise label is used as a system input, and speech features are mapped to high level representation through neural network, consequently, different types of phones are expected to be differentiated. The training process introduces Siamese network and optimizes the distance between output embeddings according to whether two embeddings are from same type of phones or not, and optimization process is realized by well-adapted loss function. Results show that accuracy of Siamese network based on cosine hinge loss function which achieve accuracy of 89.93% is better than the other methods in the experiment, and accuracy of diagnosis is 89.19% in pronunciation error verification task with the best approach under the circumstance of no mispronunciation training data being involved.

---

* **收稿日期：** 2018-6-11 **定稿日期：** 2018-7-25

基金项目：国家社科基金项目（18BYY124），语言资源高精尖创新中心项目(KYR17005)，国家语委科研项目（ZDI135-51），北京语言大学梧桐创新平台项目（中央高校基本科研业务费专项资金）（16PT05）（18YJ030004）,北京语言大学研究生创新基金(17YCX139)

作者简介：王振宇（1991-），男，硕士，主要研究方向：计算机辅助语音习得，语音识别；解焱陆（1980-），男，博士、副教授，主要研究方向：计算机辅助语音习得、语音信号处理；张劲松（1968-），男，博士、教授，主要研究方向：语音习得、韵律建模、语音识别、实验语音学、计算机辅助发音教学。

Key words: mispronunciations verification; phone embedding; Siamese networks

1 引言

汉语二语学习者难以习得标准发音，即使有很多对话经验的高级汉语学习者也难以掌握正确的汉语发音和声调。计算机辅助发音教学作为有限传统教育资源的有力补充，能给予二语学习者更及时有效的帮助和反馈。计算机辅助发音训练作为计算机辅助发音教学系统的重要组成部分，在系统构建过程起着重要作用。

在之前的研究中，自动语音识别系统被应用于在音段层级的发音偏误检测任务中来评估学习者发音的正确与否，以音素为单位计算对数后验概率分数[1]来检测发音偏误[2]。Witt 和 Young [3] 引入基于概率的发音良好度方法，此方法给出的是一个归一化的对数似然比分数并在[4, 5, 6]中用于句子确认。后来出现了一些发音良好度的变体[7,8,9]，也都是基于每一个音素相对于母语者置信分数均值来设置阈值从而判断偏误。以上系统提供的音段层级的反馈是比较有指导性和直观的评价结果。

由于基于发音良好度的方法的一个重要组成部分是依赖于大量人工标注自动语音识别系统。因此，我们想探究使用弱监督的方法去获得一个有区分性的特征表示，此方法也比较适合于一些资源稀缺的情况[10,11]。之前的一些研究使用了一种叫孪生网络[12]的结构，此网络将一对标明相同与否的词对输入到两个权值共享的深度神经网络来得到话者和音段信息[13]。Synnaeve 等人根据所给数据标签类型改进了损失函数，在音素错误率上得到了和全监督方法近乎相等的结果[14]。使用声学词向量的词区分任务也已经在几个其他的研究中得以应用[15,16,17]，通过比较词向量的距离计算平均错误率，以此来衡量系统准确性。Herman 等人比较了几种用于词区分任务的方法，使用卷积孪生网络使系统得到了进一步的提升[18]。

我们的方法引入声学音素向量来确认二语学习者的发音偏误，并给出了有指导性且具体的反馈。基于前人的声学词向量想法，我们使用带有配对信息的音素，基于弱监督的方法来做音素区分任务。以定长的语音特征向量作为孪生网络的输入，判断生成的音素向量是否来源于同一音素并依此修正生成向量间的距离。结果显示使用余弦最大间隔距离损失函数的卷积孪生网络得到了最好的音素确认结果。基于此结论我们使用实验得到的最好的模型去进行二语者的发音质量评价，在不添加标注的二语发音偏误数据作为训练数据的情况下，优于基于发音良好度的方法的结果，并且模型鲁棒性也更好。

在本文中，第二部分概要描述了经典的发音良好度、DNN-HMM 方法，以及基于声学音素向量和孪生网络的音素确认的方法，第三部分对实验配置和实验过程进行具体说明，第四部分根据实验结果进行分析讨论，第五部分给出总括性的结论。

2 音素确认方法概要

该章节介绍了传统的发音评价方法-发音良好度，和经典的基于 DNN-HMM 语音识别框架的发音偏误检测的基本原理。基于对传统方法原理的思考，我们提出了用基于音素结合孪生网络的方法进行发音偏误确认。

2．1 发音良好度打分

在发音评分中，发音良好度 GOP(Goodness of Pronunciation)是最广泛使用的方法之一。此方法为句子中的每个音素都给出一个置信分数。音素 p 的发音良好度分数如下：

$$\mathrm{GOP(p)} = \left|Log\left(\frac{P(\mathrm{p}|\boldsymbol{O}^{\mathrm{p}})}{max_{\mathrm{q}\in\mathrm{Q}}P(\mathrm{q}|\boldsymbol{O}^{\mathrm{q}})}\right)\right|/\mathrm{NF(p)} \quad (1)$$

给定声学模型和正则文本，p 是标准单元，q 是对比单元，$\boldsymbol{O}^{\mathrm{p}}$是 NF(number of frames)帧音素 p 的输入特征。边界信息来源于强制对其结果，Q 是可能音素的集合。设置一个阈值

以确认当前单元是否是一个正确发音，高于此阈值即为正确反之错误，此阈值根据任务和训练数据不同可做相应调整。可以利用公式 1 计算任何给定的音素的对数后验概率，并称之为亚音段分数。我们在音素发音错误确认任务中使用的基线系统是发音良好度评价系统，该系统由在大规模母语者语料库[19]训练得到的神经网络三因子声学模型构成的。

2．2 DNN-HMM 框架

深度神经网络结合隐马尔可夫模型的声学模型建模框架式是现今在自动语音识别领域比较通用和流行的框架,在大规模的连续语音识别任务中的性能也远超传统混合高斯模型 GMM-HMM 混合模型。由此，将 DNN-HMM 模型引入发音偏误检测的声学模型建模阶段以期获得更好地系统检测性能,高迎明等人已经在[34]中将使用 DNN-HMM 混合模型训练得到的声学模型应用到发音偏误检测任务中，得到 88.6%的诊断正确率。DNN 深度神经网络是前馈人工神经网络，在它的输入和输出之间有多个隐藏层，每一层由多个用来保存参数的节点构成，用输入数据对一个多层的生成性模型—深层置信网络(deep belief network, DBN)进行拟合得到参数初值[35]。DNN 的输出层一般为 softmax 输出，从该层得到每一帧音频数据所对应的三音子音素的绑定状态的后验概率。已知从训练集估计得到的各绑定状态的先验概率，利用贝叶斯公式将先验概率转化为各状态的后验概率并输出,某状态 s 的输出概率如公式 2：

$$P(\boldsymbol{o}|\mathrm{s})=\frac{P(\mathrm{s}|\boldsymbol{o})}{P(\mathrm{s})}.const(\mathrm{s}) \quad (2)$$

其中，$\boldsymbol{o}$ 指每一帧的声学特征，$P(\mathrm{s})$就是绑定状态的先验概率，$P(\mathrm{s}|\boldsymbol{o})$是经过 DNN 得到的状态 s 的后验概率，$const(\mathrm{s})$是与绑定状态 s 无关的常量。得到各绑定状态的输出概率后，经过 HMM[33]算法得到相应的识别结果。整体框架示意图如图 1 摘自[36]。

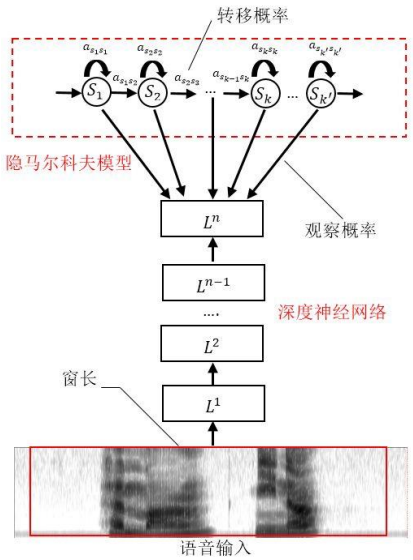

图 1：DNN-HMM 框架

2．3 带调音素向量

由于传统的发音良好度的方法的检测效果有限,而基于 DNN-HMM 语音识别框架的发音偏误检测系统需要大量标注了二语者发音偏误的数据，我们提出了音素向量的方法，期望通过得到音段层级声学特征的高层表示来区分各音素种类,从而区分二语者偏误发音和母语者标准发音。

音素区分任务将变长的语音段特征输入神经网络,神经网络最后一层的输出向量作为输入特征的高维表示，在这个向量空间中相同语音段的映射距离近，不同的类别互相远离。关键词搜索[20] 和 无监督条目搜索[21]已经使用过了类似的表示向量。在汉语中共有 60 个

音素类型，21 个声母 39 个韵母，每个汉字带一个声调（包括轻声共五类），并且声调由韵母，也就是元音来区分。在训练集中理论上应有 216 类音素类型（21+39*5），由于在汉语中部分元音不对应某些声调，其中 204 类在汉语中较为常见，所以训练集中共包括 204 类音素类型。这个分类方法期望在一个音素区分任务中同时解决确认声调和发音偏误确认两个问题。最终，不同的音素特征向量应该被映射为能有效区分音素类型的高维表示向量。

2．4 音素相似性孪生网络

这种基于配对信息的监督学习已经在一些领域中得到过应用，包括语义词向量[23,24,25]和图像方面的应用[26]。这些研究同样引入了孪生网络，该网络结构于 19 世纪 90 年代被首次提出[12]。我们的发音偏误确认任务通过判定标准发音人和二语者的发音相似性来达到评价二语者发音良好度的目的，这和孪生网络用来区分语义或者图像的方式有相似之处。孪生网络由两个权值共享的神经网络构成，首先输入两段语音特征矩阵，然后将其映射到由最后一层全连接层产生的高维向量的空间。在训练过程中，依据高维特征表示空间中的因素向量是否来自于同一类音素来调整优化他们之间的距离。在训练集中的数据标签只是配对信息而不是具体的音素标注，即每对输入特征都带有一个标签来说明他们是不是一类数据。这种辅助信息在缺乏资源或者数据稀疏的场合更容易获得，之前有研究使用无监督的条目发现系统来找未定义的匹配词对[27， 28]。

在我们的实验中，语料依据强监督的音素识别系统给出强制对齐结果切分成音素段，并且音素边界准确率在 96.26%误差在 50 毫秒。由于所有的语音数据都是文本已知的朗读语料，依据强制对齐结果得到每个音素的边界，再结合文本中音素序列给每一个音段打上对应的标签，最后根据音素类别标签生成配对信息。由于训练语料[19]中均为发音状况良好的母语者，我们默认将母语者发音作为标准音来训练模型。所以在数据标签获取过程中无需人工标注数据。图 2 描述了我们的网络结构。

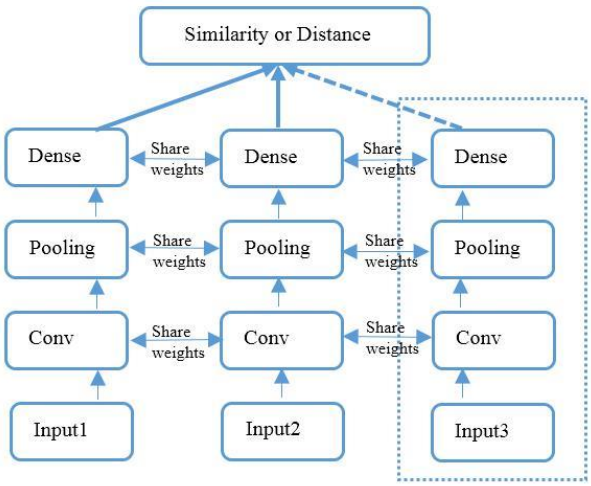

图 2：孪生网络结构（双生/三生）

图 2 展示的是以两个输入和以三个输入开始的孪生网络结构，两种输入模式对应不同的损失函数。基于欧式距离[26]（见公式 3）的损失函数更易于理解也符合网络的设计初衷，但是它更倾向于解决区分不同配对的问题，对于相同的配对效果不佳。然而，余弦距离相似性[14](见公式 4)的损失函数可以计算向量间的夹角而不再是空间距离。余弦距离相似性损失函数的最好情况是相同的向量夹角趋近于 0，不同的向量夹角趋近于正交。

$$Loss_{euc}(\boldsymbol{x_1},\boldsymbol{x_2}) = \begin{cases} euc(\boldsymbol{x_1},\boldsymbol{x_2})\ if\ same \\ 1 - euc(\boldsymbol{x_1},\boldsymbol{x_2})\ if\ different \end{cases} \quad (3)$$

With euc$(\boldsymbol{x_1}, \boldsymbol{x_2})$ = $\frac{\|\boldsymbol{x_1}-\boldsymbol{x_2}\|}{\|\boldsymbol{x_1}\|\|\boldsymbol{x_2}\|}$

$$Loss_{cos}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \begin{cases} \frac{1-\cos(\boldsymbol{x_1},\boldsymbol{x_2})}{2} \ if\ same \\ cos^2(\boldsymbol{x_1}, \boldsymbol{x_2}) \ if\ different \end{cases} \quad (4)$$

With $\cos(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle}{\|\boldsymbol{x_1}\|\|\boldsymbol{x_2}\|}$

由于我们希望将训练集中每一类和其他类区分开，并且对于多个不同的类的相似程度也不同，相对距离更适合作为损失函数中的距离衡量，并且我们假设没有在训练集中出现的配对为不同的对。由此我们引入了余弦最大间隔距离损失[24]（见公式 5）这个损失函数。

$$Loss_{\cos hinge} = \max\{0, m + d_{cos}(\boldsymbol{x_1}, \boldsymbol{x_2}) - d(\boldsymbol{x_1}, \boldsymbol{x_3})\} \quad （5）$$

在这里$d_{cos}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{1-\cos(\boldsymbol{x_1},\boldsymbol{x_2})}{2}$是$\boldsymbol{x_1}$和$\boldsymbol{x_2}$之间的余弦距离，m 是一个超参数-最大间隔，$\boldsymbol{x_1}$和$\boldsymbol{x_2}$来自不同一个类，$\boldsymbol{x_3}$和$\boldsymbol{x_1}$是同一类。

3 实验

本章介绍了所有实验用到的实验数据，给出了基于计算音素后验概率的发音良好度方法的实验配置和部分实现细节，以及基于音素向量的孪生网络的网络配置和实验过程。

3．1 语料

863 语音识别语音语料库[19]用作训练数据，其中 10%的数据用作开发集数据。测试数据分两部分，不同实验目的下使用不同测试语料。用母语者数据测试模型的性能，用二语者数据来做发音偏误确认实验。所有测试语料来自北京语言大学中介语语音语料库[29]。数据描述如下：

表 1：测试集数据

小时数	约 13 小时
话者	7 二语女性，6 母语男性，6 母语女性
句子数	5469
音素数	81142
平均句子长度	14 音节

表 2：训练集数据

小时数	约 110 小时
话者	836 母语男性，83 母语女性
句子数	96745
音素数	2351095
平均句子长度	12 音节

3．2 发音良好度评价系统

我们使用 kaldi 语音识别工具箱[30]实现发音良好度评价系统，训练出一个上下文相关的 HMM-DNN 声学模型，基于声学模型输出的后验概率为每个音素给定一个音段层级的发音分

数。使用 48 维声学特征，包括 13 维 MFCC 和 3 维音高还有各自的一阶和二阶差分系数。深度神经网络包括六个全连接层，每一层有 1024 个单元。输出层使用 softmax 函数产生 2943 个帧级别音素概率状态类型。输入为 11 帧向量，由当前帧和前后五帧拼接而成。给定强制对齐结果，使用发音良好度评价系统得到的帧级别的对数后验概率分数，通过公式 1 计算发音良好度分数，设置阈值为 0.5 来给出一个这个音是否发对的二择一判断。结果表明，发音良好度系统在母语者数据上的测试结果能达到 86.32%准确率。

3．3 基于音素向量的评价系统

提取特征阶段以 10ms 为帧移 20ms 为窗长提取 MFCC 特征和音高以及各自的一阶和二阶差分系数，共 48 维声学特征。声学音素向量的方法要求将定长的语音特征向量映射到定长的特征表示空间中。由此我们将帧数较长的音素段利用动态时间规整[22]方法，将帧数较短的音素段使用补零的方法，统一归整 18 帧，即 0.018 秒。动态时间规整的方法的缺点之一就是需要计算大量的对齐距离，而且不管是动态时间规整还是补零对原始信息都有一定程度的损失和扭曲，结合两个方法的目的也是为了最大程度上缓解原始信息的扭曲。同时，对每句话做全局均值方差归一化[31]以尽量消除话者或者其他方面信息的干扰。

我们使用了大约 100 小时的母语者数据来做音素对，整个训练数据包括开发集产生 235 万个音素段，这些数据被分批加入到孪生网络中训练。每批数据有 512 个条目，可产生六万个音素对，我们随机挑选其中 3 万对，并且相同对和不相同对各半，以保证训练数据平衡。测试分两步，先用母语者数据测试以检测模型的性能，然后使用二语者数据在性能最好的模型上做音素区分实验，并与发音良好度评价系统结果进行比较。所有测试数据文本来源一致。

3. 4 孪生网络配置

我们使用利用 tensorflow 作为后台的 keras 工具包实现孪生网络。使用 ADADELTA[32]作为随机优化方法，ADADELTA 的优点是依照过去梯度的积累来调整学习率。网络结构描述如下：

DNN SIA：2048 个节点的全连接层，激活函数 RELU；1024 个节点的全连接层，激活函数 RELU；256 个节点的全连接层，激活函数为线性激活函数

CNN SIA：96 个过滤器的一维卷积层对每 9 帧进行过滤，激活函数 RELU 最大池化层，步长为 3；96 个过滤器的一维卷积层对每 8 帧进行过滤，激活函数 RELU 最大池化层，步长为 3；1024 个节点的全连接层，激活函数为 RELU；256 节点的全连接层，线性激活函数，损失函数是基于欧氏距离的损失函数或者是基于余弦相似性的损失函数。

CNN TRI:和 CNN SIA 的结构相同，只是网络被复制成了三份，接受三个输入，损失函数余弦最大间隔损失函数。

我们比较了不同类型的损失函数和网络结构，最后使用余弦最大间隔距离损失函数的三输入孪生网络达到了最好的效果，边界参数 m 为 0.15。

3. 5 评价指标

对于母语者数据我们以预测结果是否和根据标注文本得到的配对信息相对应来衡量模型的精度。对于二语者数据，基于实验中的四种情况：接受率，拒绝率，错误接受率，错误拒绝率。最后该实验包括三个指标来评价偏误确认系统的性能，分别是：

False Rejection Rate(FRR):正确的发音被诊断为错误发音的数量占全部正确发音的比例。

False Acceptance Rate(FAR):错误的发音被诊断为正确的数量占全部错误发音的数量占全部的比例。

Diagnostic Accuracy:预测结果和标签一致的比例，即正确的被诊断为正确的，偏误发音被预测为偏误的比例。

4 结果

表 3 描述了在母语测试数据上的模型准确率结果。每个模型的阈值都是 0.5，设置成 0.5 的原因是，针对根据声韵母标注而来的配对标签，每次预测的过程其实都是二分类问题，因为随机的概率是 0.5，所以每个模型给出的预测概率必须大于 0.5 才算预测正确。

表 3：在母语者数据集上的测试结果

方法	准确率
GOP	86.32%
DNN SIA(euc loss)	83.37%
CNN SIA（euc loss）	86.46%
CNN SIA (cos sim loss)	87.83%
CNN TRI (hinge cos loss)	89.93%

分析以上结果我们发现，使用余弦最大间隔损失的三输入的孪生网络达到最好的效果。高迎明等人结合了一些词典扩展和特征融合的技巧[34]使用二语者数据训练基于 DNN-HMM 框架的语音识别系统，来进行发音偏误监测任务。我们在相同的测试数据集上，用表三中所有的方法训练得到的模型来进行音素发音偏误确认的实验，并与发音良好度模型和[34]中的 DNN-HMM 模型进行对比，DNN-HMM 的结果来自于高迎明的实验结果[34]。结果如表 4：

表 4：在二语者数据集上的测试结果

方法	**FRR**	**FAR**	**准确率**
GOP	23.86%	32.81%	74.67%
DNN SIA(euc loss)	10.48%	43.29%	82.61%
CNN SIA(euc loss)	8.44%	37.89%	85.52%
CNN SIA(cos sim loss)	7.31%	34.88%	87.13%
DNN-HMM	5.5%	34.6%	88.60%
CNN TRI (hinge cos loss)	5.4%	32.48%	89.19%

从结果中我们发现，发音良好度的方法的效果下降非常明显，原因是训练数据和测试数据的不匹配造成的，训练数据为母语者，而测试数据为二语者的话，二语者产生的非标准音素发音被当作未知音素，其识别结果无法估计，由此造成了较大损失。而 DNN-HMM 模型效果较好的原因是训练数据和测试数据匹配程度较高，而且基于强监督学习方法依赖音段层级的人工标注。相比之下孪生网络的方法就有较好的鲁棒性和可实践性，原因是结合孪生网络的区分原则，如果是在训练过程中没有出现的配对就视为是不同的，那么恰好二语者发音的非标准音素发音就被视为了和标准音不同的类型。并且我们知道相比母语者数据，二语者数据更难以收集，所以孪生网络训练音素向量的方法也有更好地可行性。这里我们还尝试了调整孪生网络最后一层生成的向量的维度，结果表明最后一层维度是 128 维的情况下，音素区分正确率最高，如图 3。

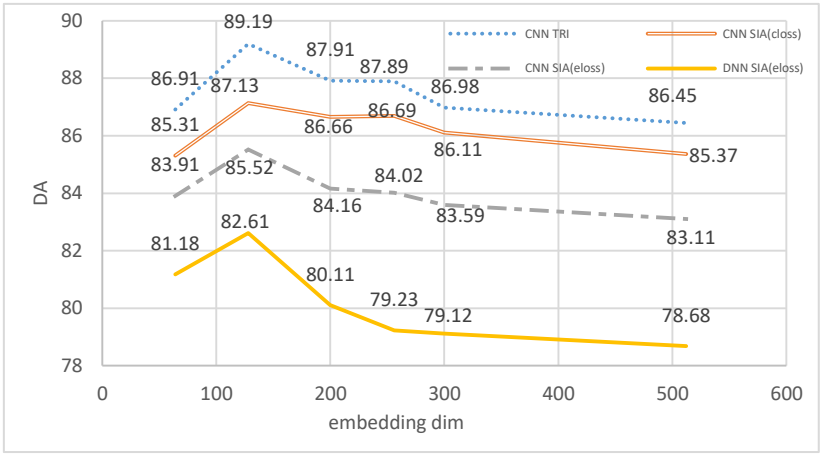

图 3：调参结果

图 4 展示了我们的方法应用在音素发音偏误确认任务的一个例子，发音偏误确认系统给出了该句中每个音素与标准音的相似度分数，该条数据来自于二语者数据。该句是一个日本女性发音人的音频数据，内容为“很忙，你呢”。通过人工听辨发现，其中“很”和“呢”的发音有较明显的声调错误。

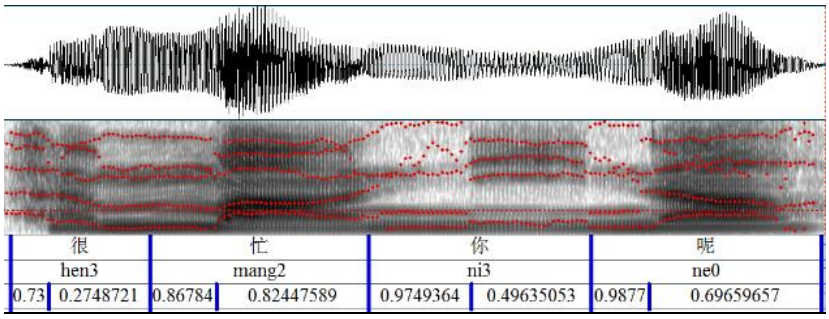

图 4：确认样例

5. 结论

发音评估是二语教学中比较重要的一个环节，而传统教学方法难以及时和有针对性的给出二语学习者有效的帮助和反馈，所以我们希望用计算机辅助发音教学来弥补传统教学方法的不足。其中计算机辅助发音训练是影响计算机辅助发音教学系统性能的重要部分。由于二语者的数据和母语标准模板数据在听觉感知上有较明显差异，结合音素向量可以作为输入特征的高层特征表示和孪生网络能够区分输入特征向量的相似性的特点，本文中我们提出了一种基于声学音素向量和孪生网络的方法来训练音素区分模型，之后依据系统给出的二语者和母语者的发音相似程度来给二语者的发音提供一个音素层级的评估打分。二语者可以根据该有指导性意义的打分来提高自己的发音水平。对比发音良好度基线系统和基于 DNN-HMM 框架的偏误检测系统，我们的方法训练得到的模型的鲁棒性更优而且训练数据以及相应标签也更易获得，在音素诊断诊断正确率上达到 89.19%的效果。

参考文献

[1] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,” in Proc. Eurospeech, 1999.

[2] Franco H, Neumeyer L, Ramos M, et al. Automatic Detection of Phone-Level Mispronunciation for Language Learning[C]// European Conference on Speech Communication and Technology, Eurospeech 1999, Budapest, Hungary, September. DBLP, 1999:851--854.

[3] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech Communication, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] Lee, Chin Hui, and R. A. Sukkar. "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition.", US 5675706 A. 1997.

[5] M. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative Utterance Verification for Connected Digit Recognition," IEEE Trans. on Speech and Audio Proc., Vol. 5, No. 3, pp. 266-277, May 1997.

[6] Sukkar, Rafid A., et al. "Verifying and correcting recognition string hypotheses using discriminative utterance verification." Speech Communication 22.4(1997):333-342.

[7] Zheng, Jing, et al. "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation." IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2007: IV-201 - IV-204.

[8] S. Wei, G. Hu, Y. Hu, R.H. Wang, “A new method for mispronunciation detection using support vector machine based on pronunciation space models,” Speech Communications, vol. 51, no. 10, pp. 896–905, 2009.

[9] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, “Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers,” Speech Communication, 67, pp. 154- 166, 2015.

[10] Park, Alex S., and J. R. Glass. "Unsupervised Pattern Discovery in Speech." IEEE Transactions on Audio Speech & Language Processing 16.1(2008):186-197.

[11] Jansen, Aren, K. Church, and H. Hermansky. "Towards spoken term discovery at scale with zero resources." INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September DBLP, 2010:1676-1679.

[12] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. S¨ackinger, and R. Shah, “Signature verification using a “siamese” time delay neural network,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 7, no. 04,pp. 669–688, 1993.

[13] Zeghidour, Neil, et al. "Joint Learning of Speaker and Phonetic Similarities with Siamese Networks." INTERSPEECH 2016:1295-1299.

[14] Synnaeve, Gabriel, T. Schatz, and E. Dupoux. "Phonetics embedding learning with side information." Spoken Language Technology Workshop IEEE, 2014:106-111.

[15] A. Jansen, S. Thomas, and H. Hermansky, “Weak top-down constraints for unsupervised acoustic model training,” in Proc. ICASSP, 2013.

[16] H. Kamper, M. Elsner, A. Jansen, and S. J. Goldwater, “Unsupervised neural network based feature extraction using weak top-down constraints,” in Proc. ICASSP, 2015.

[17] R. Thiolli`ere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, “A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling,” in Proc. INTERSPEECH, 2015.

[18] Kamper, Herman, W. Wang, and K. Livescu. "Deep convolutional acoustic word embeddings using word-pair side information." IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2016:4950-4954.

[19] Xu, Bo, et al. "Update Progress of Sinohear: Advanced Mandarin LVCSR System at NLPR." The Proceedings of the 2000.

[20] K. Levin, A. Jansen, and B. Van Durme, “Segmental acoustic indexing for zero resource keyword search,” in Proc. ICASSP, 2015.

[21] H. Kamper, A. Jansen, and S. Goldwater, “Fully unsupervised small-vocabulary speech recognition using a segmental Bayesian model,” in Proc. INTERSPEECH, 2015.

[22] Sakoe, Hiroaki, and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition." IEEE Transactions on Acoustics Speech & Signal Processing 26.1(2003):43-49.

[23] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck,“Learning deep structured semantic models for web search using clickthrough data,” in Proc. CIMK, 2013.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv: 1301.3781, 2013.

[25] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, “From paraphrase database to compositional paraphrase model and back,” Trans. ACL, vol. 3, pp. 345–358, 2015.

[26] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in Proc. CVPR, 2006.

[27] A. S. Park and J. R. Glass, “Unsupervised pattern discovery in speech,” IEEE Trans. Audio, Speech, Language Process., vol.16, no. 1, pp. 186–197, 2008.

[28] A. Jansen and B. Van Durme, “Efficient spoken term discovery using randomized algorithms,” in Proc. ASRU, 2011.

[29] W. Cao, D. Wang J. Zhang, and Z. Xiong, "Developing a Chinese L2 speech database of Japanese learners with narrow phonetic labels for computer assisted pronunciation training," INTERSPEECH, 2010.

[30] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." Idiap (2012).

[31] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” Speech Communication, vol. 25, no. 1, pp. 133–147, 1998.

[32] M. D. Zeiler, “ADADELTA: An adaptive learning rate method,” arXiv preprint arXiv: 1212.5701, 2012.

[33] Rabiner, L. R. "A tutorial on hidden Markov models and selected applications in speech recognition." Readings in Speech Recognition 77.2(1990):267-296.

[34] Gao, Yingming, et al. "A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network." INTERSPEECH 2016: Conference of the International Speech Communication Association 2015.

[35] Hinton, Geoffrey E., S. Osindero, and Y. W. Teh. "A Fast Learning Algorithm for Deep Belief Nets." Neural Computation 18.7(2006):1527-1554.

[36] 俞栋, 邓力. 解析深度学习:语音识别实践[M]. 电子工业出版社, 2016.

作者联系方式：姓名：王振宇　地址：北京市海淀区学院路 15 号北京语言大学 邮编：100083
电话： 13161756988 电子邮箱：m18535229387@163.com