Journal of Computer Applications
计算机应用,2014,34(6):1694 - 1698
ISSN 1001-9081
CODEN JYIIDU
2014-06-10
http://www.joca.cn

文章编号:1001-9081(2014)06-1694-05
doi:10.11772/j.issn.1001-9081.2014.06.1694

发音错误检测中基于多数据流的 Tandem 特征方法

袁 桦[1,2*],蔡 猛[1,2],赵军红[3,4,5],张卫强[1,2],刘 加[1,2]
(1.清华大学 电子工程系,北京 100084; 2.清华信息科学与技术国家实验室(清华大学),北京 100084;
3.中国科学院 电子学研究所,北京 100190; 4.传感技术国家重点实验室(中国科学院),北京 100190;
5.中国科学院大学,北京 100190)
(*通信作者电子邮箱 yuanh08@mails.tsinghua.edu.cn)

摘 要:针对发音错误检测中标注的发音数据资源有限的情况,提出在 Tandem 系统框架下利用其他数据来提高特征的区分性。以中国人的英语发音为研究对象,选取了相对容易获取的无校正发音数据、母语普通话和母语英语作为辅助数据,实验结果表明,这几种数据都能够有效地提高系统性能,其中无校正数据表现出最好的性能。同时,比较了不同的扩展帧长,以多层神经感知(MLP)和深度神经网络(DNN)作为典型的浅层和深层神经网络,以及 Tandem 特征的不同结构对系统性能的影响。最后,多数据流融合的策略用于进一步提高系统性能,基于 DNN 的无校正发音数据流和母语英语数据流合并的 Tandem 特征取得了最好的性能,与基线系统相比,识别正确率提高了 7.96%,错误类型诊断正确率提高了 14.71%。

关键词:发音错误检测;Tandem 特征;发音规则;深度神经网络(DNN);多层神经感知(MLP)
中图分类号:TP391.42 **文献标志码**:A

Multi-stream based Tandem feature method for mispronunciation detection

YUAN Hua[1,2*], CAI Meng[1,2], ZHAO Junhong[3,4,5], ZHANG Weiqiang[1,2], LIU Jia[1,2]
(1. *Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*;
2. *Tsinghua National Laboratory for Information Science and Technology (Tsinghua University), Beijing* 100084, *China*;
3. *Institute of Electronics, Chinese Academy of Sciences, Beijing* 100190, *China*;
4. *State Key Laboratory of Transducer Technology (Chinese Academy of Sciences), Beijing* 100190, *China*;
5. *University of Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: To deal with the under-resourced labeled pronunciation data in mispronunciation detection, some other data were used to improve the discriminability of feature in the framework of Tandem system. Taking Chinese learning of English as object, unlabeled data, native Mandarin data and native English data which can be relatively easily accessed were selected as the assisted data. The experiments show that these types of data can effectively improve the performance of system, and the unlabeled data performs the best. And the effect to system performance was discussed with different length of frame context, the shallow and deep neural network typically represented by Multi-Layer Perception (MLP) and Deep Neural Network (DNN), and different structure of Tandem feature. Finally the strategy of merging multiple data streams was used to further improve the system performance, and the best system performance was achieved by combining the DNN based unlabeled data stream and native English stream. Compared with the baseline system, the recognition accuracy is increased by 7.96%, and the diagnostic accuracy of mispronunciation type is increased by 14.71%.

Key words: mispronunciation detection; Tandem feature; phonological rule; Deep Neural Network (DNN); Multi-Layer Perception (MLP)

0 引言

计算机辅助语言学习(Computer-Assisted Language Learning,CALL)系统是计算机相关技术发展下的产物,因能够给语言学习者提供一个生动有趣且方便独立的学习环境而被广泛使用。发音错误检测技术是新时代 CALL 系统中的关键技术,其目标是检测出学习者发音中的单个发音错误,以便于后端反馈系统能够提供具体的错误信息和改正方法,从根本上帮助学习者提高发音水平。

发音错误检测技术起源于文献[1],并得到了 Yoon 等[2-4]的关注,主要方法是利用置信度分数来判断当前发音是否为正确发音。随着研究的深入,研究人员开始探索如何检测出具体的发音错误类型(将发音 A 错误成发音 B)[5-10],这样后端系统可以提供更精细的错误信息给学习者。这种检测方法都是首先构建包含正确发音和错误发音的预测性检测网络,然后使用声学模型和检测器得到学习者的实际发音。在检测网络不变的情况下,声学模型的区分性决定了最终检测结果的准确性。

收稿日期:2013-12-16;**修回日期**:2014-01-30。 **基金项目**:国家自然科学基金资助项目(61370034,61273268,61005019,61105017)。
作者简介:袁桦(1985 -),女,湖北浠水人,博士研究生,主要研究方向:发音错误检测; 蔡猛(1987 -),男,河北沧州人,博士研究生,主要研究方向:自动语音识别; 赵军红(1987 -),女,山东菏泽人,博士研究生,主要研究方向:语音合成; 张卫强(1979 -),男,河北雄县人,助理研究员,博士,主要研究方向:模式识别; 刘加(1954 -),男,福建福州人,教授,博士,主要研究方向:语音信号处理。

而在发音错误检测中,学习者发音数据的采集和专家对发音的音素级别的精细标注需要大量的成本。在目前的研究报道中,非母语学习的发音数据库规模都非常有限,文献[7-10]中所采集的数据库都只包含了 10~20 h 语音。训练数据的不足会对声学模型的区分性造成很大影响。但是在目前的发音错误检测中,这方面的研究非常欠缺。文献[8-9]引入语音识别中的区分性训练方法来代替传统的最大似然训练方法,文献[10]首次报道了深度神经网络(Deep Neural Network, DNN)在发音错误检测中的性能,并在预训练中使用无标注的发音数据来提高性能。这些方法都从一定程度上提高了模型的性能,但是区分性的训练方法并不能解决数据受限的问题,在 DNN 的预训练中使用辅助数据的作用也是非常有限的,这些方法都不能很好地解决训练数据受限的问题。

由于人工神经网络(Artificial Neural Network, ANN)具有良好的区分性,本文采用基于 ANN 的 Tandem 系统框架,针对发音学习数据库难于采集和标注的现状,尝试利用其他辅助数据来提高声学模型的区分性。虽然经过人工标注的非母语英语数据是有限的,但是标准母语英语的数据资源非常丰富。同时,对于中国人说英语而言,母语普通话的数据资源也容易获得。此外,没有经过人工标注的发音学习数据带有原始的参考文本,经数据统计表明,学习者的实际发音与参考文本中的标注一般都具有 80% 的一致性。ANN 的区分性结构有助于从这些无校正的数据中提取出有用的区分性信息。本文构建了一种基于多数据流融合的系统框架,利用这几种不同性质甚至是不同语言的数据来提高系统性能,最后通过实验验证了方法的有效性。

1　发音错误检测系统

1.1　发音规则的提取

发音错误预测所针对的规律性错误,一般都是由“语言迁移”所引起[11],学习者将自己母语中的发音习惯方式等引入到第二语言的学习当中,当两种发音不同时就会造成发音错误。一般用发音规则对这种发音错误进行描述:

$$\partial \rightarrow \beta / \lambda\_\omega \tag{1}$$

表示位于发音 λ 和发音 ω 中间的发音 ∂,会被错误发音为 β。∂ 和 β 可以表示相连的单个或者多个发音,λ 和 ω 则限制为单个发音。表示插入错误或者删除错误时则分别将 ∂ 或者 β 取值为空。

本文采取数据驱动的方法进行规则提取。首先从数据库中提取出标准发音音素串和对应的人工校正音素串,再用动态规划算法对这两个音素串进行对齐,从中找出不匹配的部分,以发音规则的形式表示。

1.2　发音错误检测系统框架

发音错误检测系统的处理流程如图 1 所示。学习者按照系统提示的文本进行发音后,系统根据该文本结合标准发音字典和发音规则生成音素级别的检测网络。从学习者的语音中提取出声学特征后,利用发音检测器结合声学模型和检测网络进行识别,得到学习者最终的实际发音音素串。再将该音素串与文本所对应的标准音素串进行对准,得到发音错误的音素以及错误类型。将这些检错结果输入到后端的错误反馈中,就可以结合具体的错误情况,给予学习者不同的提示信息。

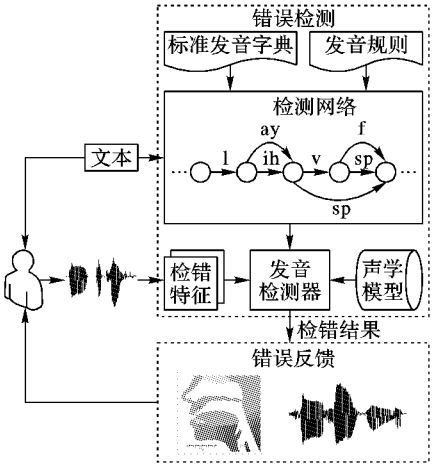

图 1　发音错误检测系统处理流程

2　基于多数据流的 Tandem 特征

ANN 是一种仿造生物神经网络的统计数据建模工具,因具有良好的记忆能力和非线性映射等优点,在语音识别等人工智能领域得到广泛应用。目前比较典型的神经网络有多层神经感知器(Multi-Layer Perception, MLP)和 DNN[12-13]。MLP 是多层前馈网络,由一层输入层、一层或多层隐含层和一层输出层组成,模型学习都是通过误差反向传播算法实现的。DNN 的结构与 MLP 的结构类似,但是在训练开始前会使用无监督的方法进行初始化,这样多层次的神经网络也能够得到很好的训练。具有多层隐含层的 DNN 使得语音识别的正确率得到了巨大的提升。

将 ANN 输出的后验概率取对数、去相关和降维后,作为一种新的“特征”来替代传统的感知线性预测(Perception Linear Predictive, PLP)系数等特征,用于训练声学模型和解码,这种方法被称作 Tandem 方法[14-15],其中的新特征被称作 Tandem 特征。Tandem 方法方便与其他声学模型的区分性方法相结合,因此得到广泛应用。Tandem 系统借助于 ANN,将声学特征的提取过程由传统的固定提取模式转变为可以训练的过程,使得特征能够根据相应的任务变得更具有区分性,同时也方便于利用交叉域或者交叉语言的数据来提高系统性能[16-18]。本文尝试在 Tandem 系统框架的基础上,从与学习者的英语发音数据相关联的母语英语数据、母语普通话数据和无校正的英语发音数据中提取出有用的信息,并将这些信息结合起来提高特征的区分性。神经网络结构的差异会对性能造成影响,本文对浅层网络和深层网络也进行了对比分析。下面首先对 MLP 和 DNN 进行简单介绍,再详细描述 Tandem 特征的提取方法。

2.1　神经网络 MLP 和 DNN

MLP 隐含层使用 Sigmoid 激活函数,输出层使用 Softmax 函数,其相应的代价函数为目标概率 d 和 Softmax 输出的概率 p 之间的交叉熵:

$$c = -\sum_j d_j \log p_j \tag{2}$$

其中 j 表示输出层单元的索引。本文中使用的浅层神经网络是典型的三层的前向神经网络 MLP,由 1 层输入层、1 层隐含层和 1 层输出层组成,其中隐含层包含了 2 048 个节点。

DNN 多层次的参数能够更好地捕捉到语音数据的高阶统计结构。为了较好地训练 DNN 中多层次的网络,在训练之前对 DNN 的参数进行无监督的逐层初始化(Layer-wise Pre-training),也称为预训练,由低层向高层,每次更新一层参数。初始化过程中,一共使用了两种受限玻尔兹曼机(Restricted Boltzmann Machine, RBM):1)高斯伯努利 RBM 用于描述底层假设服从高斯分布的输入特征;2)伯努利 RBM 用于对高层的二进制数据进行有效的描述。初始化完成之后,DNN 与一般的前向神经网络一样,使用反向传播算法进行更新。这个训练阶段需要使用各帧语音特征的标签,以使得每类之间具有区分性。在使用反向传播算法对 DNN 进行训练时,输出层同样使用的是 Softmax 函数。本文中使用的深层神经网络是包含了 5 层隐含层的 DNN,每个隐含层同样包含了 2 048 个节点。

2.2 Tandem 特征

Tandem 系统中可以包含多个级联的神经网络,一般会包含一到两级。每个神经网络输出的 Tandem 特征,可以直接输入到高斯混合模型(Gaussian Mixture Model, GMM)—隐含马尔可夫模型(Hidden Markov Model, HMM)系统中,也可以继续输入到新的神经网络中再进行处理。本文中所使用的 Tandem 特征处理如图 2 所示,包含了单级和级联 Tandem 特征的处理。神经网络输出的后验概率取对数后经主成分分析(Principal Component Analysis, PCA)处理降维成为 Tandem 特征。第一级的神经网络可以使用目标数据、交叉域或者交叉语言的数据进行训练。这样目标数据经过第一级训练好的神经网络处理后,就进行了从滤波器组系数(Filterbank, Fbank)到目标数据、交叉域或者交叉语言的音素后验概率的转换。对于第二级神经网络,本文只使用目标数据进行训练,使得到的后验概率与目标数据的特性更加匹配。

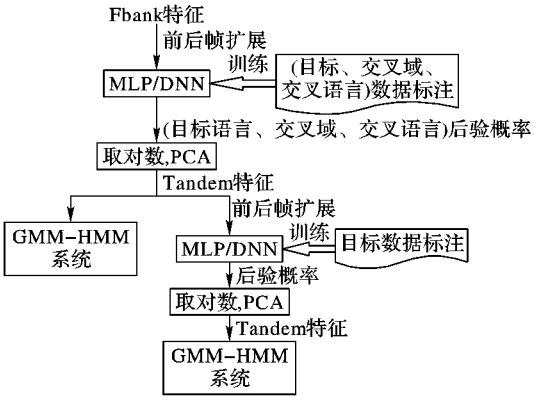

图 2 单级和级联 Tandem 特征处理流程

2.3 基于多数据流的 Tandem 特征

在级联 Tandem 特征的基础上,本文提出如图 3 所示的基于多数据流的 Tandem 特征。第一级的神经网络分别使用无校正的发音数据、母语普通话发音数据和母语英语进行训练。这些数据在不同的方面与目标数据具有相似的特点,由这些数据分别训练的神经网络输出的后验概率会存在某种程度的互补性。将这些后验概率进行特征组合进一步输入后到第二级的神经网络中,使用目标数据进行训练,最后将得到的 Tandem 特征输入到 GMM-HMM 系统中。这样第一级不同神经网络输出的信息就在第二级神经网络中得到融合。

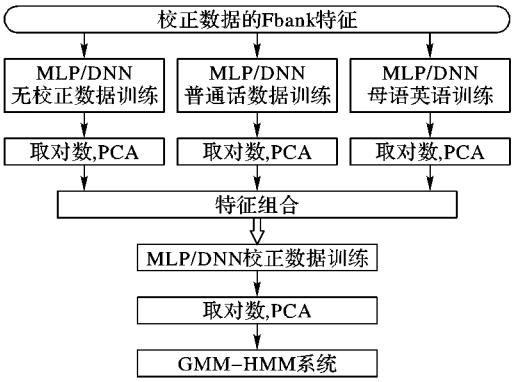

图 3 基于多数据流的 Tandem 特征提取流程

3 实验配置

3.1 数据库

本文实验采用了 CU-CHLOE 语言学习数据库、WSJ1 数据库和 863 普通话数据库。CU-CHLOE 数据库中包含了 111 人中国人的英语语音,每人的录音文本为 86 句英语教师设计的语句和 237 句 TIMIT 文本。其中英语教师设计的语句录音都由经过培训的语言学专家进行了音素级别的标注,而 TIMIT 文本所对应的语音由于资源匮乏没有进行标注。本文从已标注的数据中随机选择 25 名女性和 30 名男性的语音作为训练集,其余的作为测试集。WSJ1 数据库和 863 普通话数据库分别作为母语英语和母语普通话使用。WSJ1 训练集中包含了 245 人的语音,863 普通话数据库中包含了 126 人的语音。各数据库的数据量分布情况如表 1 所示。

表 1 数据量统计

数据类别	数据量/h
CHLOE 训练集(labeled)	9.6
CHLOE 测试集	10.0
CHLOE 无校正数据(unlabeled)	50.4
普通话数据库(Mandarin)	90.9
WSJ1 数据库(WSJ)	145.6

3.2 参数设置

从 CHLOE 训练集上一共提取出 3 506 条发音规则用于错误检测。基线模型使用了 13 维的 PLP 特征以及一阶和二阶差分。在决策树聚类后,模型中包含了 1 076 个三音子状态,每个状态用 8 个高斯混合分量来描述。神经网络的训练使用经过了谱均值方差归一化的 40 维 Fbank 特征和 1 维能量。所有 Tandem 特征均降维为 39 维,与 PLP 特征的维数相等。神经网络在训练中所使用的标签是各数据集的 PLP-GMM-HMM 模型对准得到的单音子状态。

3.3 评价指标

本文采用了 4 个性能评价指标:音素识别正确率 *CORR*、错误类型正确率 *DA*、错误接受率 *FAR* 和错误拒绝率 *FRR*,它们的计算方法为:

$$\begin{cases} CORR = \dfrac{N_{\text{reccorrect}}}{N_{\text{all}}} \\ DA = \dfrac{N_{\text{hit}}}{N_{\text{TR}} + N_{\text{FR}}} \\ FAR = \dfrac{N_{\text{FA}}}{N_{\text{incorrect}}} \\ FRR = \dfrac{N_{\text{FR}}}{N_{\text{correct}}} \end{cases} \tag{3}$$

其中:N_{all} 表示所有的音素个数,N_{correct} 表示实际发音正确的音素个数,$N_{\text{incorrect}}$ 表示实际发音错误的音素个数,$N_{\text{reccorrect}}$ 表示识别正确的音素个数,N_{hit} 表示检测出的错误类型与实际的错误类型一致的音素个数,N_{TR} 表示实际的错误发音被检测为错误发音的个数,N_{FR} 表示实际的正确发音被检测为错误发音的个数,N_{FA} 表示实际的错误发音被检测为正确发音的个数。*FAR* 和 *FRR* 值越小越好,*DA* 和 *CORR* 值越大越好。当错误接受、错误拒绝和错误诊断类型这三类错误最少时,*CORR* 取得最大值。

4　实验结果与分析

4.1　基线系统性能

本文首先对 PLP 基线模型进行测试。为了能与 Tandem 区分性特征进行对比,在基线模型的基础上采用最小音素错误率准则(Minimum Phone Error, MPE)进行训练,对得到的 MPE 模型也进行了测试,测试结果如表 2 所示。在进行 MPE 训练后,系统性能有了较大的提升。

表 2　基线系统的性能　　%

模型	*CORR*	*DA*	*FAR*	*FRR*
PLP 基线系统	74.49	27.79	34.24	20.64
PLP(MPE)	78.74	34.72	33.86	15.55

4.2　不同扩展帧数和不同数据对 Tandem 特征的影响

在 Tandem 特征的实验中,首先测试不同扩展帧数对性能的影响。图 4 给出了基于 MLP 的 Tandem 特征在扩展帧数为 11,21 和 31 时的 *CORR* 性能。

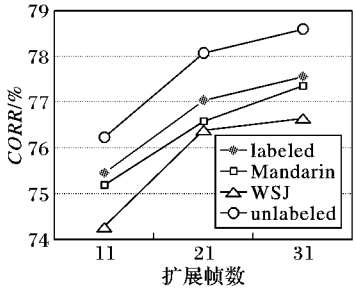

图 4　基于 MLP 的 Tandem 特征在不同扩展帧数下的 *CORR* 性能

可以看到,第 31 帧时 Tandem 特征所取得的性能是最好的,这与各数据库的特点是符合的。因为这几个数据库中所包含的均为朗读发音(WSJ 中包含少量的自然发音),朗读者为了使每个单词发音清楚,速度偏慢,单个发音的时间比一般的口语对话中的发音时间更长,所以较长的扩展帧数能够对发音更好地建模。对不同扩展帧数时的性能进行对比,随着扩展帧数的增长,提升的幅度下降,说明增大扩展帧数对性能的提升是有限的,随着扩展帧数的增大,区分信息越多,但引入的发音之间的混淆性也会增大。扩展帧数并不是越大越好,31 帧的取值对于本文的任务是比较合适的。下面的实验均使用扩展帧数为 31 帧。

对不同数据训练得到的 Tandem 特征进行对比,可以发现 unlabeled 的性能是最好的,然后依次是 labeled、Mandarin 和 WSJ。这是因为 unlabeled 数据与 CHLOE 中的标注数据是最匹配的,虽然其中包含了部分错误发音,但是 MLP 能够从大量的数据中学习到各发音的主要区分信息。Mandarin 和 WSJ 则均与 CHLOE 的数据存在比较大的失配。

4.3　级联神经网络结构对 Tandem 特征的影响

在单级 Tandem 特征的基础上,对级联 Tandem 特征进行测试,神经网络使用 MLP,结果如图 5 所示,级联的 Tandem 特征相比单级 Tandem 特征有较大的性能提升。因为级联 Tandem 特征能够在第一级 MLP 的基础上,使用更长的上下文信息,同时还能从第一级的输出中选择最有区分性的互补信息。在级联 Tandem 特征中,两级 MLP 都使用 labeled 数据训练对性能的提升是最小的,因为这种情况下可互补的信息是最小的,而使用了其他辅助数据的 MLP 系统都能比没有使用辅助数据的系统取得更好的结果。WSJ 和 Mandarin 分别作为不同域和不同语言的数据,在这种级联结构下都能起到有利的辅助作用。

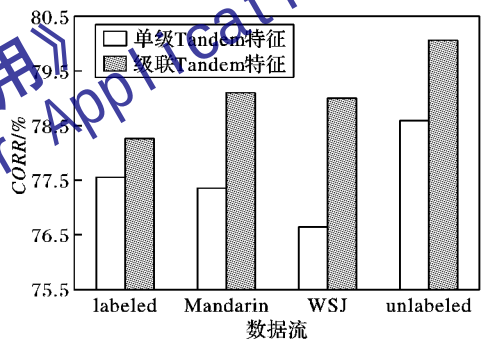

图 5　不同数据流下单级和级联 Tandem 特征的 *CORR* 性能

4.4　浅层网络和深层网络的对比

上面都是 Tandem 特征基于浅层网络 MLP 的性能。本文在此基础上,进一步分析深层网络 DNN 下不同数据的性能,并将它们的结果进行对比。图 6 是不同网络结构下级联 Tandem 特征的 *CORR* 性能。DNN 使整体的性能都得到提升,unlabeled 数据流的性能提升最显著。这再次验证了 unlabeled 数据的有效性。与文献[10]中将 unlabeled 数据用于辅助 DNN 的预训练相比,这种级联的形式可以更有效地利用 unlabeled 数据中标签 80% 的正确性,能够取得更好的效果。

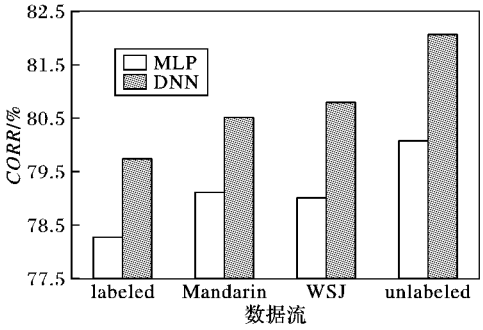

图 6　不同网络结构下级联 Tandem 特征的 CORR 性能

4.5　多数据流融合的策略

对不同神经网络输出的数据流进行合并,可以使不同数

据间的区分信息进行互补。因此本文进一步的对各数据流进行融合,并得到实验结果如表 3 所示。数据流之间两两进行合并,都能在各自单数据流的基础上使性能得到提升。其中性能最好的是 unlabeled 数据流和 Mandarin 数据流进行合并,取得的 *CORR* 为 82.45%,相比基线系统的 74.49%,绝对提高了 7.96%。而最终的三个数据流合并没有比该数据流合并取得更好的性能。这说明三个数据流合并的提升空间是有限的,虽然数据流越多,能从其中提取的区分性信息越多,但是同时也可能引入混杂的信息,并且使网络结构变得更加复杂。

表 3 基于 DNN 的多数据流合并的级联 Tandem 特征 %

数据流	*CORR*	*DA*	*FAR*	*FRR*
Mandarin & WSJ	81.59	40.58	33.47	12.62
unlabeled & WSJ	82.45	42.50	34.06	11.54
unlabeled & Mandarin	82.21	41.91	34.11	11.80
Mandarin & WSJ & unlabeled	82.35	42.24	33.95	11.67

5 结语

声学模型在发音错误检测中对区分正确和错误发音起到重要作用,本文在 Tandem 特征的基础上针对有标注的发音学习数据受限的问题借助其他容易获取的关联数据来提高各发音声学模型之间的区分性。本文对不同的扩展帧长、不同数据、不同结构和不同层次的神经网络对 Tandem 特征造成的影响都进行了研究分析。最后借助基于多数据流的 Tandem 系统框架,挖掘不同数据流之间的互补性。unlabeled 数据流和 Mandarin 数据流融合取得最好的系统性能,与基线系统相比,*CORR* 提高了 7.96%,*DA* 提高了 14.71%,*FAR* 降低了 0.18%,*FRR* 降低了 9.10%。

参考文献:

[1] FRANCO H, NEUMEYER L, RAMOS M, *et al*. Automatic detection of phone-level mispronunciation for language learning [EB/OL]. [2013-10-10]. http://www.speech.sri.com/people/hef/papers/F020.PS.

[2] YOON S Y, HASEGAWA-JOHNSON M, SPROAT R. Landmark-based automated pronunciation error detection [EB/OL]. [2013-10-10]. http://www.isle.illinois.edu/sst/pubs/2010/yoon10interspeech.pdf.

[3] WEI S, HU G, HU Y, *et al*. A new method for mispronunciation detection using support vector machine based on pronunciation space models [J]. Speech Communication, 2009, 51(10): 896 – 905.

[4] LI H, HUANG S, WANG S, *et al*. Context-dependent duration modeling with backoff strategy and look-up tables for pronunciation assessment and mispronunciation detection [C]// Proceedings of the 12th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2011: 1133 – 1136.

[5] HARRISON A M, LO W K, QIAN X, *et al*. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training [C]// Proceedings of the 2009 Speech and Language Technology in Education Workshop. Baixas: ISCA, 2009: 137 – 140.

[6] LO W K, ZHANG S, MENG H. Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system [C]// Proceedings of the 11th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2010: 765 – 768.

[7] WANG Y B, LEE L S. Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training [C]// Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2012: 5049 – 5052.

[8] STANLEY T, HACIOGLU K. Improving L1-specific phonological error diagnosis in computer assisted pronunciation training [C]// Proceedings of the 13th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2012: 826 – 829.

[9] QIAN X, SOONG F K, MENG H M. Discriminative acoustic model for improving mispronunciation detection and diagnosis in Computer-Aided Pronunciation Training (CAPT) [C]// Proceedings of the 11th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2010: 757 – 760.

[10] QIAN X, MENG H M, SOONG F K. The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 english to support computer-aided pronunciation training [C]// Proceedings of the 13th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2012: 774 – 777.

[11] GASS S M, SELINKER L. Language learning in language transfer [M]. Philadelphia: John Benjamins Publishing Company, 1993: 87 – 101.

[12] DAHL G E, YU D, DENG L, *et al*. Context - dependent pre - trained deep neural networks for large-vocabulary speech recognition [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30 – 42.

[13] MOHAMED A, DAHL G E, HINTON G. Acoustic modeling using deep belief networks [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14 – 22.

[14] HERMANSKY H, ELLIS D P W, SHARMA S. Tandem connectionist feature extraction for conventional HMM systems [C]// Proceedings of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2000: 1635 – 1638.

[15] ZHENG X, WU Z, SHEN B, *et al*. Investigation of tandem deep belief network approach for phoneme recognition [C]// Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2013: 7586 – 7590.

[16] QIAN Y, LIU J. Articulatory feature based multilingual MLPs for low-resource speech recognition [C]// Proceedings of the 13th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2012, 3: 2601 – 2604.

[17] QIAN Y, LIU J. Cross-lingual and ensemble MLPs strategies for low-resource speech recognition [C]// Proceedings of the 13th Annual Conference of the International Speech Communication Association. Baixas: ISCA, 2012, 3: 2581 – 2584.

[18] TUSKE Z, PINTO J, WILLETT D, *et al*. Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions [C]// Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2013: 7349 – 7353.