北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.029

基于发音特征的发音偏误趋势检测研究

屈乐园 解焱陆† 张劲松

北京语言大学信息科学学院, 北京 100083; † 通信作者, E-mail: xieyanlu@blcu.edu.cn

摘要 为了提升计算机辅助发音训练(CAPT)系统中发音偏误趋势(PET)的检测效果, 确保反馈信息的准确性与有效性, 提出一种基于对数似然比的发音特征方法。该方法将多个基于深度神经网络的发音特征提取器用于生成帧级别的对数似然比, 然后将对数似然比组成的发音特征用于PET的检测, 为学习者提供发音位置和发音方法的正音信息。实验结果表明, 发音特征对PET的检测效果优于常用声学特征(MFCC, PLP和fBank), 当发音特征与MFCC特征相结合时, 可以进一步提升性能, 达到错误接受率为5.0%, 错误拒绝率为30.8%, 诊断正确率为89.8%的检测效果。

关键词 发音特征; 发音偏误趋势; 计算机辅助发音训练; 对数似然比

中图分类号 TP391; H193

A Study of Articulatory Features Based Detection of Pronunciation Erroneous Tendency

QU Leyuan, XIE Yanlu†, ZHANG Jinsong

School of Information Science, Beijing Language and Culture University, Beijing 100083;
† Corresponding author, E-mail: xieyanlu@blcu.edu.cn

Abstract This paper proposed to apply senone log-likelihood ratio based articulatory features (AFs) to improve pronunciation erroneous tendency (PET) detection performance. The feedback information of articulation-placement and articulation-manner could be derived from the definition of PET. The framework of the method involved two main steps. 1) A bank of attribute extractors based on neural networks were trained to estimate the log-likelihood ratio (LLR) for each senone at a frame level. 2) AFs composed of those LLRs outputted from each attribute extractor were used for detecting PETs. Results demonstrated that the proposed system had better performance than the baseline system using MFCC. Moreover, substantial improvements were obtained by combining AFs with MFCC, achieving a lower false rejection rate of 5.0%, a lower false acceptance rate of 30.8% and a higher diagnostic accuracy of 89.8%.

Key words articulatory features (AFs); pronunciation erroneous tendency (PET); computer assisted pronunciation training (CAPT); senone log-likelihood ratios

近年来, 随着计算机硬件和深度学习的发展, 计算机辅助发音训练(CAPT)成为当前研究热点之一。CAPT 作为计算机辅助语言学习系统的重要组成部分, 可以有效地加强二语学习者的口语能力, 因此备受语音识别、语言学和教育学等领域学者关注[1]。

正音反馈信息对学习者有重要的意义[2], 但在 CAPT 系统中, 学习者无法辨识自己的错误发音。Neri 等[3]发现, 即使以有限的形式实现正音反馈信息, 也能改善学习者在音素层级的发音质量, 同时对学习者的学习动力也有积极作用。基于自动语音识别(ASR)技术为学习者提供反馈信息的方法

北京语言大学梧桐创新平台项目(16PT05)和北京语言大学研究生创新基金项目(16YCX160)资助
收稿日期: 2016–07–29; 修回日期: 2016–10–01; 网络出版日期: 2016–11–30

可分为两类。一类是基于置信分数度量的方法，这类方法通过计算置信分数来衡量标准发音与二语者发音之间的差异，比如对数后验概率方法[4]、发音良好度方法(GOP)[5]以及基于 GOP 的改进方法[6–8]。这类方法计算简单，可以直接利用语音识别的中间结果，但只能为学习者提供分数上的反馈，当学习者面对一个低分数时，却不知道如何纠正自己的发音[9]。另一类是基于规则的方法，这类方法通过对比不同语言之间的差异，构建发音扩展词典[10]，或者通过统计语料库得出错误发音规则和相应频次，再用先验概率拓展发音词典[11]。与基于置信分数度量的方法相比，基于规则的方法可以为学习者提供更多音素层级的反馈信息，比如系统检测出学习者将“ret”(/r E t/) 发音为“let”(/l E t/)，就可以提示“你将/r/发成了/l/”。

上述基于规则的发音错误检测方法将学习者的错误发音归为音素替换。但是，音素替换的错误往往发生在初级学习者身上，而中高级水平学习者的偏误发音并不是简单的音素插入、删除和替换，而是相对标准发音的少许偏离，即偏误发音往往介于两个音位类型之间，而非绝对的音位替换[12]。因此，Cao 等[13]根据发音位置和发音方法的不准确性，定义了相应的发音偏误趋势，包括高化、低化、前化、后化等 64 种。在检测中，将发音偏误趋势加入扩展发音网络中，不仅可以检测出学习者的偏误发音，而且可直接为学习者提供发音位置和发音方法的反馈。例如，学习者在练习发圆唇 u 时，容易发生展唇化偏误(将圆唇 u 误发成了展唇 u{w})，系统会提示学生“发 u 时嘴唇稍微圆一些”。Duan 等[14]和 Gao 等[15]通过对比不同模型和不同声学参数对发音偏误趋势的检测效果，验证了该方法的可行性。

发音偏误趋势是从发音位置与发音方法的角度定义的，而发音特征可以用来检测发音器官的变化。因此，本文利用发音特征改善发音偏误趋势的检测效果，为学习者提供更加详细可靠的正音反馈信息。将发音特征引入 CAPT 系统中，需要实现发音特征的提取。提取方式有 3 种：1) 利用 X 光射线仪透视或微型线圈采集说话人的发音运动信息，获取发音特征；2) 使用逆滤波的方式，对语音信号进行加工处理，提取发音特征[16]；3) 使用概率统计方法，建立相应的数学模型，将每一帧物理信号转换为不同的发音特征，用对数似然比或后验概率表示发音特征[17]。在实际操作中，语音语料数量庞大，对每位发音人都使用 X 光射线仪透视或微型线圈采集发音特征的可行性低；当前逆滤波器的还原精度不高，会直接影响偏误发音的检测效果。近年来，深度学习技术快速发展，广泛应用到语音识别、图像识别等领域，并取得良好效果。因此，本文使用概率统计方法，借助深度学习技术，实现发音特征的鲁棒提取。

1 发音偏误趋势

发音偏误趋势是相对于标准发音的少许偏离，其声学表现与正确发音十分相似。目前发音偏误检测常用的声学特征主要是频谱或倒谱特征，例如基于人耳听觉感知频率敏感曲线的梅尔频率倒谱特征(MFCC)以及在此基础之上的梅尔倒谱感知线性预测系数(MFPLP)等。这类特征对环境变化比较敏感，在不同的声学环境中，检测性能会有所不同[17]；同时，对说话人之间的差异也比较敏感，不同的发音器官、不同的说话风格等也会导致偏误检测系统性能的变化[18]。而且，这些传统的频谱或倒谱特征对声学上相似的发音区分能力较弱。因此，如何选取更具区分性的特征，有力地刻画偏误发音与正确发音之间的细微差别，完成 PET 的准确检测，是本文关注的重点。

二语学习者在学习汉语时，由于受到母语负迁移等作用的影响，倾向于使用母语中相似音的发音位置和发音方法来代替二语中的发音位置和方法。如果二语中的发音位置或方法在母语中不存在，学习者将很难正确掌握二语中的发音。Cao 等[13]根据二语学习者发音位置和发音方法的不准确性定义了相应的发音偏误趋势，部分 PET 标注符号及标注规范如表 1 所示。

表 1 PET 标注规范[14]

Table 1 PET annotation convention[14]

类型	标注符号	偏误举例	说明
舌页化	sh	sh{sh}	sh 被发成舌页音
圆唇化	o	e{o}	展唇音被发成圆唇音
后化	–	n{–}	前鼻音近似成后鼻音
短化	;	p{;}	p 送气时长较短

2　发音特征

发音特征(articulatory features, AFs)是语音产生过程中对发音器官主要动作属性的描述, 通过发音特征能够建立语音信号与主要发音单元之间的对应关系[19]。相对于一般声学特征(频谱或倒谱特征), 发音特征有诸多优势。首先, 发音特征可以描述发音器官的变化情况, 为协同发音的分析和音素序列的恢复提供更多潜在的信息, 而声学分析却不能完整而精确地揭示协同发音深层次的成因[20−21]; 其次, 发音特征独立于声学环境的变化, 可以很好地解决说话者频谱差异、背景噪音以及室内混响等问题。例如, 当发一个圆唇元音时, 所有共振峰都将向低频偏移, 这样的变化并不会因为说话者口腔形状的不同或背景噪音的干扰而发生变化[20,22]。已经有一些将发音特征应用到自动语音识别中的研究, 并取得良好效果[22−24]。

2.1　发音特征类别

汉语普通话音节可以分为声母和韵母两个部分。声母发音特征可以从发音位置和发音方法的角度来划分。韵母按口型可以分为撮口呼、齐齿呼、合口呼和开口呼, 按结构可以分为单元音韵母和复合韵母。具体的发音特征与音素的对应关系见文献[25]。实验使用的发音特征见表 2。对表 2 列出的特征分别建立相应的提取器, 用于发音特征的提取。

2.2　发音特征的提取

2.2.1　Senone 的定义

在连续语流中, 由于受到上下文语境影响, 音段的声学表现与孤立音节的情形十分不同。针对该情况, 通常使用以音素为单位的上下文相关建模方法。一些音素对于上下文音素的影响是相似的, 所以可以通过音素解码后的状态聚类进行区分, 聚类的结果称为 Senone。借鉴当前语音识别的经验, 我们在建立特征提取器时, 以基于语境信息的发音特征 Senone 为基本单元, 使用表 1 列出的转换规则, 将基于语境的音素 Senone 转换为特征 Senone(例如, 将 n–i+h 转换为鼻音–单元音+擦音), 并分别对每种发音特征建立特定的提取器, 以确保发音特征提取的准确性。

2.2.2　Senone 对数似然比

以上述定义的特征 Senone 为基本单元, 为每类发音特征分别建立一个含有 N 个 Senone 单元的特征提取器, 每一个 Senone 单元都可以用一个含有 S 个状态的模型表示。在 t 时刻, 每一个 Senone 单元内的每一个状态 $s(1\leqslant s\leqslant S)$ 的声学后验概率 $p(i,\ s|t)$都可以由解码器直接得到。每一个 Senone 单元在 t 时刻的声学后验概率可以通过与其对应的所有状态的后验概率的加和得到:

$$p(i\,|\,t)=\sum_{\forall s}p(i,s\,|\,t),\quad i=1,...,N\ 。\tag{1}$$

将 $p(i|t)$作为似然值, 把待计算 Senone 的先验概率设为 0.5, 将剩余 0.5 平均分配给其他 Senone, 因此, 第 t 帧对应的对数似然比可由式(2)计算得到:

$$\mathrm{LLR}(i\,|\,t)=\log\frac{p(i\,|\,t)}{\dfrac{1}{N-1}\sum_{\forall j\neq i}p(j\,|\,t)},\ i=1,...,N\ 。\tag{2}$$

2.2.3　基于 Senone 对数似然比的发音特征

如图 1 所示, 将语音帧分别输入每个发音特征提取器中, 根据提取器的输出以及式(1)和(2), 计算得到帧级别的对数似然比。这些对数似然比经过特征融合模块, 最后生成发音特征。发音特征的维数等于 4 个发音特征提取器 Senone 的总个数, 由此得到的特征维数将会非常庞大。因此, 我们使用线性判别分析(PCA), 对发音特征进行降维, 降维后的特征将用于发音偏误趋势的检测。

3　系统描述

本文中使用自动语音识别框架与扩展发音网络

表 2　发音特征列表

Table 2　Articulatory attributes

音段	类别	发音特征
声母	发音方法	塞音、塞擦音、擦音、鼻音、边音、送气、清音、浊音
	发音位置	双唇音、唇齿音、舌尖前、舌尖中、舌尖后、舌面前、舌面后
韵母	按口型分	撮口呼、齐齿呼、合口呼、开口呼
	按结构分	单元音、双元音、三元音、前鼻音韵母、后鼻音韵母

相结合的方法，来实现发音偏误趋势的自动检测功能。系统检测框架如图 2 所示，具体流程描述如下。

1) 系统提示学习者要读的学习文本，同时，系统根据学习文本产生相应的扩展发音网络，如图 3 所示，扩展发音网络是对学习者所有可能发音的一种表示形式(图 3 中括号内为偏误标注信息)。

2) 将学习者的发音送入发音特征提取器，并提取发音特征。

3) 使用发音特征进行声学模型的匹配。

4) 对比识别出的音素序列和标准发音序列，做出系统决策。

5) 根据发音偏误知识库，给出学习者偏误发音的纠正方法。

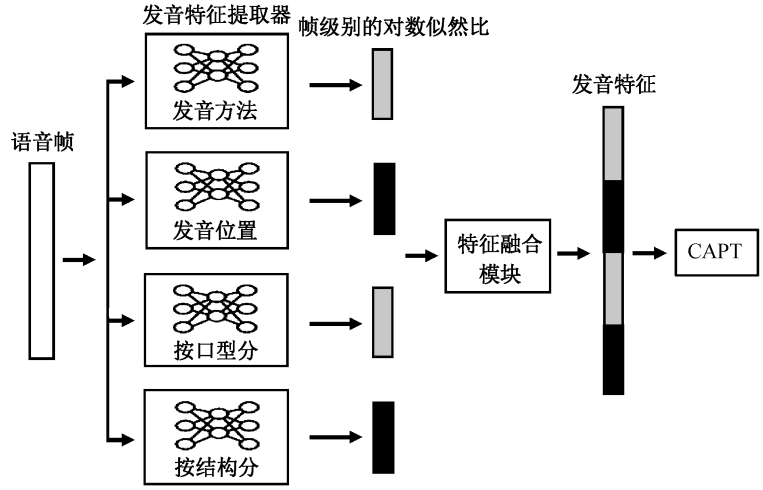

图 1 发音特征提取流程

Fig. 1 Flow chart of attribute extractor

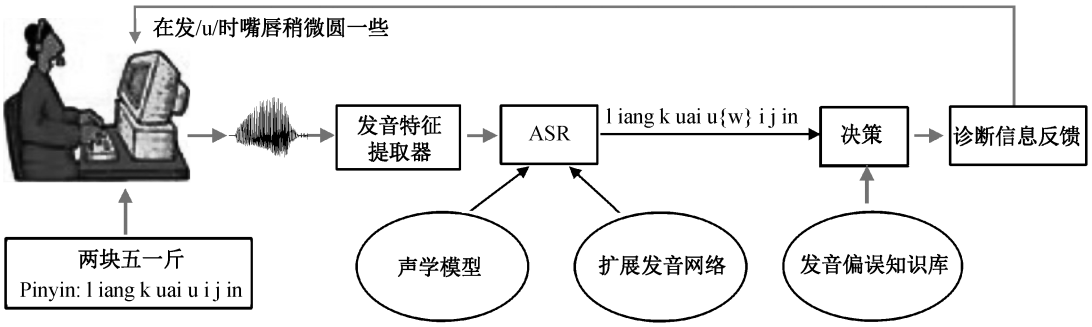

图 2 系统检测框架

Fig. 2 Overview of proposed detection framework

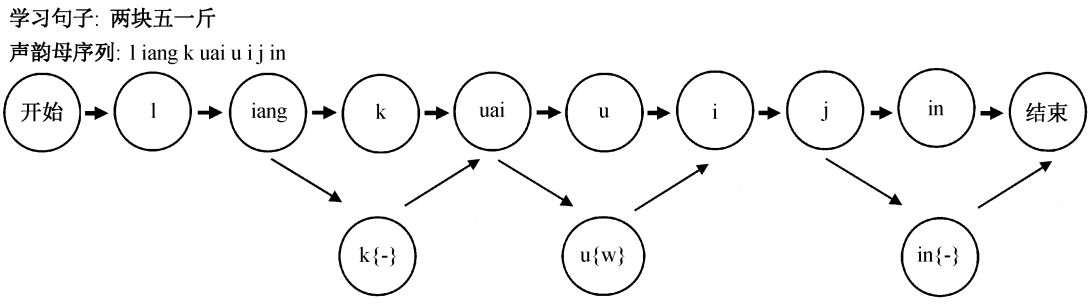

图 3 扩展发音网络

Fig. 3 Extended pronunciation network

4　实验

4.1　实验语料

4.1.1　母语者语料库

母语者语料库来自文献[26]。该语料库为大词汇量连续汉语语音识别任务而设计, 由 116 个发音者, 93271 句话, 约 100 小时的语料组成。母语者语料库将用于发音特征提取器的建立以及发音特征的提取。

4.1.2　汉语中介语语料库

中介语语料来自北京语言大学中介语语音语料库[13]。本实验所用语料取自其中 7 位日本女性发音者的连续语音, 每人约 301 句话(日常用语)。6 位语音学专业的研究生对其进行交叉标注。当出现不一致时, 请语音学专家进行判定。实验语料统计结果如表 3 所示。中介语语料库将用于发音偏误检测系统的建立。

为了避免由于某些偏误样本数量较少, 造成偏误模型训练不充分的问题, 我们只考虑前 16 种偏误发音的检测情况。16 种偏误发音可以分为以下 4 类, 最终结果也会按照这 4 类呈现。

1) 唇形圆展偏误: 发圆唇音时唇形有些展化或发展唇音时嘴唇出现圆化现象。

2) 舌位前后或鼻音前后偏误类型: 发元音时舌位过于靠前或靠后, 前鼻音近似发成后鼻音。

3) 短化偏误类型: 发送气音时送气时长较短。

4) 舌叶化偏误类型: 发舌尖后音或舌面前音时, 错发成舌叶音。

4.2　评价指标

实验的检测结果共有 4 种: 正确接受(TA)、正确拒绝(TR)、错误接受(FA)和错误拒绝(FR), 如表 4 所示。

根据这 4 种检测结果, 可以计算出以下 3 种常见的评价指标。

表 3　中介语语音语料库

Table 3　A L2 inter-Chinese corpus

语料内容	描述
文本	301 句日常用语
发音人数	7 个日本女生
句子总数	1899
音素总数	26431
每句话平均音素数	14
标注者人数	6
每句话标注者人数	2

表 4　4 种检测结果

Table 4　4 outcomes used in the experiment

结果	描述
TA	正确发音检测为正确发音的个数
TR	偏误发音检测为偏误发音的个数
FA	偏误发音检测为正确发音的个数
FR	正确发音检测为偏误发音的个数

1) 错误接受率(FAR): 学习者的错误发音被检测为正确发音的百分比。

2) 错误拒绝率(FRR): 学习者的正确发音被检测为错误发音的百分比。

3) 诊断正确率(DA): 正确发音被检测为正确, 错误发音被检测为错误的百分比。

$$\mathrm{FAR}=\frac{\mathrm{FA}}{\mathrm{FA}+\mathrm{TR}}, \tag{3}$$

$$\mathrm{FRR}=\frac{\mathrm{FR}}{\mathrm{FR}+\mathrm{TA}}, \tag{4}$$

$$\mathrm{DA}=\frac{\mathrm{TA}+\mathrm{TR}}{\mathrm{TA}+\mathrm{TR}+\mathrm{FA}+\mathrm{FR}}。 \tag{5}$$

4.3　实验配置和试验结果

4.3.1　发音特征提取器

使用深度神经网络模型(DNN)为表 1 中每一类发音特征类别分别建立一个发音特征提取器。声学特征使用 13 维的 MFCC 特征及其一阶和二阶差分, 共 39 维。实际输入到 DNN 模型中的是相邻 11 帧(当前帧及前 5 帧和后 5 帧)拼接得到的高维向量。

训练特征提取器的过程中, 使用的每一帧对应的 Senone 标签由事先训练好的 GMM-HMM 模型通过强制对齐得到, 再通过发音特征与音素的对应关系, 将音素 Senone 转换成发音特征 Senone, 作为特征提取器的训练标签。经过比较不同的隐层数(3, 4, 5, 6, 7)和节点数(512, 1024, 2048), 最后 6 个隐层和每层 1024 个节点的 DNN 取得最好效果。使用基于 Senone 对数似然比的方法提取发音特征, 用做发音偏误趋势检测。每种发音特征帧层级的提取正确率在表 5 中列出。

从表 5 可知, 基于声母发音特征的提取正确率好于基于韵母的发音特征。原因主要是因为声母结构比韵母简单, 只由辅音构成, 而韵母由元音或者元音和辅音构成, 元音和辅音在声学信号上差异较

表 5 帧层级的发音特征提取正确率

Table 5 Classification accuracies at a frame level for the attributes

发音特征	正确率/%	发音特征	正确率/%
促音	85.3	舌尖后音	86.4
塞擦音	83.7	舌面前	80.4
送气	91.1	舌面后	89.3
浊音	78.4	开口呼	69.7
清音	92.5	齐齿呼	70.4
擦音	78.8	合口呼	63.4
鼻音	94.4	撮口呼	68.7
边音	76.1	单元音	74.1
双唇音	90.9	双元音	68.5
唇齿音	73.0	三元音	62.1
舌尖音	88.6	鼻韵母	63.9
舌尖中音	82.7		

大, 这种复杂结构对以韵母为基本单元的发音特征的提取带来困难。

4.3.2 发音偏误趋势检测结果

我们训练了两种声学模型, 用于发音偏误趋势检测。一种使用语音识别常用声学特征 MFCC 的检测系统, 也是本研究的基线系统; 另一种使用本文提出的基于发音特征训练的偏误检测系统。两种系统都使用 DNN-HMM 的混合模型。

我们将最高频的 16 种发音偏误趋势划分为 4 类: 舌页化偏误、前后化偏误、唇形偏误和短化偏误。在上述 3 种评价指标中, 我们希望在最大化诊断正确率的同时, 尽量降低两类错误率。但是, 错误拒绝率和错误接受率之间相互制约。考虑到 CAPT 的目的是避免将学习者的正确发音判别为偏误发音, 因此, 我们以最大化诊断正确率和最小化错误拒绝率为目标, 进行参数优化。

图 4 为 MFCC 和 AFs(发音特征)系统对 4 类偏误的检测效果。从图 4 可以看到, 发音特征对 4 类偏误的检测效果都优于 MFCC 特征, 验证了发音特征更能描述发音器官的变化情况, 并对声学信号上比较相似的音段具有较好的区分效果。

为了进一步验证发音特征对发音偏误趋势检测的有效性, 我们将结果与文献[15]进行对比。文献[15]中对比了 3 种不同声学特征(MFCC, PLP 和 fBank)对发音偏误趋势的检测效果, 最后通过系统联合得到整体最佳检测效果。本研究同样验证了 MFCC 特征与发音特征相结合时, 对发音偏误趋势的检测效果(MFCC+AFs)。具体结果见表 6。

从表 6 可知, 发音特征的检测效果在 3 种评价指标上都优于其他 3 种声学特征。当 MFCC 特征

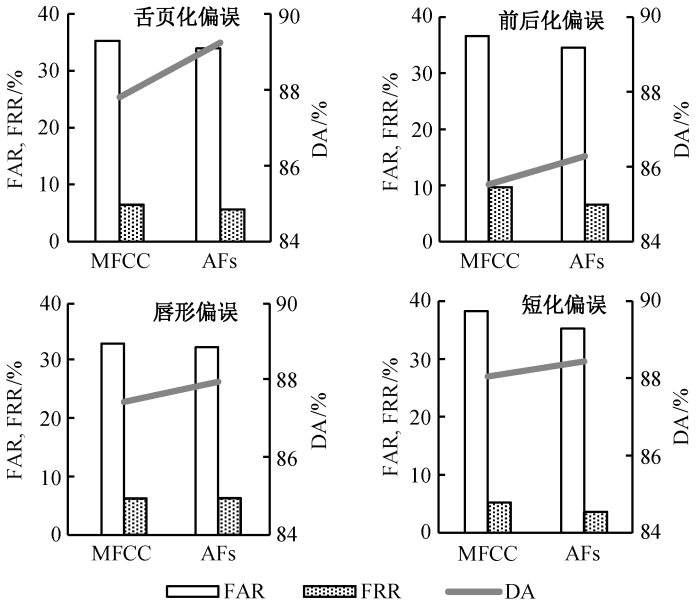

图 4 4 类偏误的检测结果

Fig. 4 Detection result of four broad heading PETs

表 6　不同方法之间的对比

Table 6　Comparison of proposed method with the other method

系统	FAR/%	FRR/%	DA/%
PLP[15]	39.4	6.1	87.4
fBank[15]	34.6	6.8	87.8
MFCC	35.2	6.5	87.9
系统融合[15]	35.6	5.5	88.6
AFs	34.7	5.8	88.3
MFCC+AFs	**30.8**	**5.0**	**89.8**

与发音特征相结合时, 达到最好检测结果, 说明发音特征在发音偏误趋势检测上, 与声学特征(MFCC)具有互补的作用, 两者的结合可以进一步提高偏误检测的正确性。

5　结语

为实现发音偏误趋势的准确检测, 本文将发音特征引入计算机辅助发音训练系统。首先, 使用深度神经网络为每一类发音特征分别建立一个特征提取器, 将提取器输出的帧层级的对数似然比作为发音特征。然后, 利用提取的发音特征, 建立发音偏误趋势检测系统。实验对比了发音特征(AFs)与 3 种常用声学特征(MFCC, PLP 和 fBank)对发音偏误的检测效果。结果表明, 基于知识的发音特征对发音偏误趋势有更好的检测效果, 在 3 个评价指标(FAR, FRR 和 DA)上都优于常用声学特征; 当发音特征与 MFCC 特征相结合时, 达到最好效果, 实现错误接受率为 5.0%, 错误拒绝率为 30.8%, 诊断正确率为 89.8%的检测结果。

参考文献

[1] Eskenazi M. An overview of spoken language technology for education. Speech Communication, 2009, 51(10): 832–844

[2] Dlaska A, Krekeler C. Self-assessment of pronunciation. System, 2008, 36(4): 506–516

[3] Neri A, Cucchiarini C, Strik H, et al. The pedagogy-technology interface in computer assisted pronunciation training. Computer Assisted Language learning, 2002, 15(5): 441–467

[4] Franco H, Neumeyer L, Ramos M, et al. Automatic detection of phone-level mispronunciation for language learning // EUROSPEECH. Budapest, 1999: 851–854

[5] Witt S M, Young S J. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 2000, 30(2): 95–108

[6] Zhang F, Huang C, Soong F K, et al. Automatic mispronunciation detection for Mandarin // 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Caesars Palace, 2008: 5077–5080

[7] Wang Y B, Lee L S. Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training // 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, 2012: 5049–5052

[8] Hu W, Qian Y, Soong F K, et al. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Communication, 2015, 67: 154–166

[9] Witt S M. Automatic error detection in pronunciation training: where we are and where we need to go // Proc ISADEPT. Stockholm, 2012: 12–19

[10] 王岚, 李崇国, 蒙美玲, 等. 音素级错误发音自动检测. 先进技术研究通报, 2009, 3(2): 6–10

[11] Luo D, Yang X, Wang L. Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus // INTERSPEECH. Florence, 2011: 1593–1596

[12] Yoon S Y, Hasegawa-Johnson M, Sproat R. Landmark-based automated pronunciation error detection // INTERSPEECH. Makuhari, 2010: 614–617

[13] Cao W, Wang D, Zhang J, et al. Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training // Eleventh Annual Conference of the International Speech Communication Association. Chiba, 2010: 1922–1925

[14] Duan R, Zhang J, Cao W, et al. A preliminary study on ASR-based detection of Chinese mispronunciation

by Japanese learners // INTERSPEECH. Singapore, 2014: 1478–1481

[15] Gao Y, Xie Y, Cao W, et al. A study on robust detection of pronunciation erroneous tendency based on deep neural network // 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, 2015: 693–696

[16] Schroeter J, Sondhi M M. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Transactions on Speech and Audio Processing, 1994, 2(1): 133–150

[17] Yu D, Siniscalchi S M, Deng L, et al. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition // 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, 2012: 4169–4172

[18] Kirchhoff K. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments // ICSLP. Sydney, 1998: 1–4

[19] Kirchho K. Robust speech recognition using articulatory information [D]. Bielefeld: University of Bielefeld, 1999

[20] 张晴晴, 潘接林, 颜永红. 基于发音特征的汉语普通话语音声学建模. 声学学报, 2010, 35(2): 254–260

[21] Metze F. Articulatory features for conversational speech recognition [D]. Karlsruhe: Karlsruhe University, 2005

[22] Livescu K, Cetin O, Hasegawa-Johnson M, et al. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop // 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Pacific Grove, CA, 2007: IV-621–IV-624

[23] Cetin O, Kantor A, King S, et al. An articulatory feature-based tandem approach and factored observation modeling // 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Pacific Grove, CA, 2007: IV-645–IV-648

[24] Çetin O, Magimai-Doss M, Livescu K, et al. Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs // IEEE Workshop on Automatic Speech Recognition & Understanding. Kyoto, 2007: 36–41

[25] 黄伯荣, 廖序东. 现代汉语. 北京: 高等教育出版社, 2007

[26] Gao S, Xu B, Zhang H, et al. Update progress of Sinohear: advanced Mandarin LVCSR system at NLPR // INTERSPEECH. Beijing, 2000: 798–801