发音属性优化建模及其在偏误检测的应用

郭铭昊
语言资源高精尖创新中心/北京
北京语言大学信息科学学院/北京
gmhgmh8000@163.com

解焱陆
语言资源高精尖创新中心/北京
北京语言大学信息科学学院/北京
xieyanlu@blcu.edu.cn

摘要

近年来，发音属性常常被用于计算机辅助发音训练系统（CAPT）中。本文针对使用发音属性的一些难点，提出了一种建模细颗粒度发音属性（FSA）的方法，并在跨语言属性识别、发音偏误检测中进行测试。最终，我们得到了最优平均识别准确率约为 95%的属性检测器组；在两个二语测试集上的偏误检测，相比基线，基于 FSA 方法均获得了超过 1%的性能提升。此外，我们还根据发音属性的跨语言特性设置了对照实验，并在上述任务中测试和分析。

关键词：发音属性；偏误检测；属性识别

Speech attributes optimization modeling and application in mispronunciation detection

Minghao Guo
Beijing Advanced Innovation Center for Language Resources/Beijing Beijing Language and Culture University/Beijing
gmhgmh8000@163.com

Yanlu Xie
Beijing Advanced Innovation Center for Language Resources/Beijing Beijing Language and Culture University/Beijing
xieyanlu@blcu.edu.cn

Abstract

In recent years, Speech attributes are often used in computer-aided pronunciation training systems (CAPT). This paper proposes a method for modeling fine-grained speech attributes (FSA) for some difficulties in using speech attributes, and tests in cross-language attribute recognition and mispronunciation detection. In the end, we obtained an attribute detector group with an optimal average recognition accuracy rate of about 95%; the mispronunciation detection on the two second language test sets, based on the FSA method achieved a performance improvement of more than 1% compared to the baseline. In addition, according to the cross-language characteristics of speech attribute, we set up a comparative experiment and tested and analyzed in the above tasks.

Keywords： Speech attribute; Mispronunciation detection;Attribute recognition

第十九届中国计算语言学大会论文集，第67页-第77页，海南，中国，2020年10月30日至11月1日。

1 引言

近年来，随着二语学习需求的增长，学习汉语的人越来越多。基于自动语音识别的计算机发音训练系统（The computer-aided pronunciation training system）不仅能够满足当下学习者碎片化学习时间的需要还能弥补传统课堂教学的劣势。它的主要核心功能有：（1）提供反馈；（2）评估发音质量。从反馈形式的角度看，CAPT 系统可大致分为发音质量打分和发音偏误检测两种类型，发音偏误检测任务的目标则是以高精度检测发音错误并给出对应的纠音反馈。研究发现，即使以简单的形式提供纠音反馈，也能够改善学习者在音素层级的发音质量（Neri A, 2006）。用于提供纠音反馈的研究有很多，例如利用拓展识别网络创建一个音素级的发音偏误检测和诊断的模型（Harrison A M, 2009），利用发音属性来提供诊断性反馈等。

通过研究人类识别语音的过程，人的记忆单元中字词存储的基本单位是段，并且通过一系列的特征集合来相互区分，这些用于描述语音学发音并区分语音段的特征称为“区分性特征”。这些特征可以从语音的不同方面定义，如发音位置、发音方式等，而这些“区分性特征”叫做发音属性（Speech Attribute）。目前，发音属性在二语学习中主要用于提供纠音反馈、简化二语语料库标注等，而发音属性的定义方法多采用国际音标标准。

外国学生在学习汉语时出现的发音偏误，往往就是由于发音位置等发音属性的不准确导致的。二语学习者受母语负迁移等作用影响，其发音属性常常会倾向于母语中相似音的发音属性，同样地，如果二语中的发音属性在其母语中缺失，则学习者将很难正确掌握新的发音方法。目前，在偏误检测任务上应用发音属性的方法有：发音偏误趋势建模、发音属性特征提取、多语言发音属性建模等。

Cao 等根据来自于发音人发音位置和发音方法等发音属性的不准确，定义了包括高化、低化、前化、后化等发音偏误趋势（Cao, 2010）。Li 等人基于发音偏误趋势的属性特征提取，用于提供诊断性反馈（Li, 2016）。但是，上述方法也存在很多局限性，例如高度依赖拥有准确标注信息的大规模二语语料库。采用多语言建模发音属性的原因在于，发音属性具备跨语言特性，且当二语者在发音时发生母语负迁移现象，其偏误发音的发音属性会包含两种语言的发音属性。因此，若同时建模两种语言的发音属性，将有助于检测偏误发音的发音属性（Duan, 2017）。理论上，通过多语言发音属性建模，有助于建模任意母语背景二语者语料的发音属性。采用多语言发音属性建模也存在难点，例如：难以建模所有已知语言、汉语与其他语言发音属性定义存在差异（如，汉语元音“i”）。以上应用中，使用发音属性的方法往往采用国际音标的定义，但是由于汉语和其他语言在发音属性的定义上存在差异，国际音标无法准确地描述汉语的发音属性。

假设在没有足够的二语数据集的情况下，本研究针对整合多母语描述发音偏误方法的难点，提出了一个以学习汉语为目的发音属性定义和优化建模方法，即细颗粒度的发音属性（FSA），将有助于改善汉语的发音偏误检测任务。在此基础上，检测属性检测器的跨语言能力，以及探究面对不同母语背景学习者语料时上述方法检测发音偏误的能力。根据发音属性具备可跨语言的特点，我们还探究了单语言训练的属性检测器的跨语言能力，通过控制建模时的上下文信息，降低了单语言属性检测器对汉语数据的过度适应，并设置了多个对照实验分别采用不同的上下文信息建模，在汉语和英语两个测试集中进行属性检测，最后对比双语言属性检测器的检测结果来进行分析。由于跨语言建模发音属性具备描述发音偏误的能力，我们还在母语为日语和俄罗斯语的学习者测试集上，进行次音段级和音段级的发音偏误检测。

2 发音属性的定义

本研究从四个方面对汉语声母进行了描述：发音位置（PA）、发音方式（MA）、是否送气（AS）、清浊音（VO）。而汉语元音部分则包括四个类别:舌位前后（TF）、舌位高低（TH）、唇形圆展（RO）、PA 和 VO。需要强调的是，在声学音标中辅音和元音的发音属性定义不同

（Siniscalchi，2008），因此我们分别对辅音和元音的发音属性进行建模，并尝试将它们在 PA 分类中合并建模。由于所有的汉语元音在 AS 和 VO 中都没有子分类，所以我们将它们的详情放在声母发音属性定义中呈现。

我们将所有的汉语辅音与 IPA 一一映射，根据 IPA 上对应音素的知识信息，找到我们需要的属性信息并给予分类标签。在 PA 类别中，所有元音部分都将被标记为“vowels”，其中声母的四个类别使用映射表 1 的音素分类中产生（C.Zhang, 2011）。在表 1 中，汉语辅音以拼音形式首先列出，其次则是以音素表示的英语辅音。该表还列出了英文中存在但中文中不存在的属性，这些属性没有参与建模，以此不难看出汉语和英语属性的区别。例如，英语中没有 AS 属性分类，以及 Timit 音素集中只有一个清化的元音“axh”。

Attributes		Phone set (Ch/En)	
PA	Bilabial	b p m	p b m w
	Labiodental	f	f v
	Alveolar	d t l n	t l el ch sh jh zh dx nx
	Dental	c s z	s dh en n r z th d
	Retroflex	zh ch sh r	
	Palatal	j q x	y
	Velar	g k h	k g ng
MA	Stop	g p d t g k	t p k b d g
	Fricative	f s sh r x h	sh th f hh dh hv v w zh s z
	Affricate	z zh c ch j q	ch jh
	Nasal	m n	en m nx ng n
	Lateral	l	el l
	Approximant		dx
	Tap or Flap		r y

表 1.中英文辅音属性类别表（部分）

汉语韵母由多个元音和鼻元音组成（en、an 等)，相对于声母来说比较复杂。因此，我们将每个汉语韵母描述为一组 IPA 音素，然后根据这些音素得到每个韵母的属性集。表 2 中列出了四个汉语元音属性类别，列出了汉语单元音和英语音素的属性分类。

此外，汉语和英语的元音在舌位上有很大的差异。在过去的研究中，将音素舌位前后大致分为三大类：前、中、后（MullerM，2017），这样简单的分类显然不能完全体现汉语元音在舌位前后的位置。为了找到更好的描述汉语的舌位前后的分类方法，我们将表示汉语元音分为五类和七类分类建模。由于五分类的舌位可以直接对应于声母的发音位置，所以我们在 PA 类中同时对韵母和声母进行建模，而在 TF 类中更详细地分为七类。将 PA 与 TF 进行比较，可以看出两种分类方法的差异，如表 2 所示。另外，汉语的声母在 TF、RO 和 TH 中被标记为“辅音”。值得注意的是，汉语韵母中存在着三种属性维度，它们描述了汉语韵母中存在的属性数量。例如，汉语的最后一个“iao”被描述为三个 IPA 音素，所以它在每个类别中都有三维属性。

Attributes		Phone set (Ch/En)	
PA	Dental	ii	
	Retroflex	iii	
	Palatal/Front	i v	iy ih ae eh
	PA-Central	a	ax ix ux axh axr er
	Velar/Back	u	aa ah ao uw uh

T H	High	i ii iii v u	ix iy ux uw
	Second H		ih uh
	Half H		
	Middle		axh axr ax
	Half L		ah ao eh er
	Second L		ae
	Low	a	aa
T F	Front 2	ii	
	Front 1	iii	
	Front	i v	ae eh iy
	Half F		ih
	Central	a	axh axr ax er ix ux
	Half B		uh uw
	Back	u	aa ah ao

表 2.中英文辅音属性类别表（部分）

3 基于 FSA 方法的优化建模

3.1 时延神经网络的设计

对连续语流数据下的语音任务来说，由于语音是一种时序序列，上下文信息对于声学模型的性能影响非常关键，在发音偏误检测任务中也是同样。TDNN 其优点在于多层网络训练时对输入特征具有较强的时序建模能力、描述了语音特征在时间序列上的关系、具备时间不变性且不需要对样本标注进行时间定位。适用于本研究的关键在于 TDNN 对动态语音分类任务具有相当好的性能表现（Waibel, 1989）。图 1 所示是本研究训练发音属性时的 TDNN 模型结构，这种 TDNN 结构对时间序列输入数据[10,11,12]具有有限的动态响应。假设 t 是当前帧，在输入层（layer1），帧[t-2，t+2]被拼接在一起。层 2,3 和 4 我们分别将帧[t-1，t+2]，[t-3，t+3]和[t-7，t+2]拼接在一起。总的来说，神经网络的左上下文为 13，右上下文为 9。

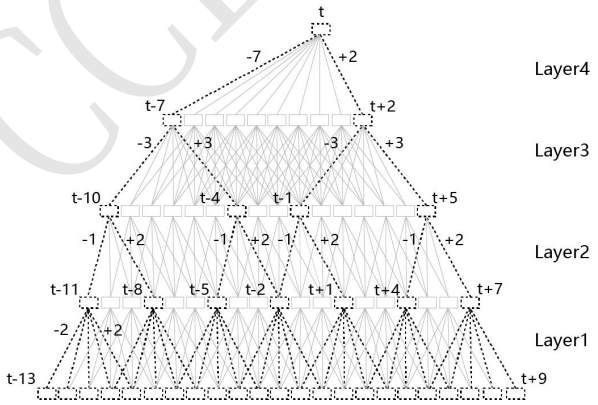


图 1.本研究的 TDNN 网络结构

3.2 I-Vector 特征的提取

我们使用所有训练集特征建立 GMM 建模通用背景模型，得到 GMM 训练的统计量后重新训练 GMM，得到 UBM，其中训练特征 40 维，高斯数 512 个；使用 UBM 初始化 i-vector，获取正规化（CMVN）的特征后验概率，计算统计量，根据统计量计算最后的 i-vector 模型 F，其中 s 维度为 512*40，m 维度 512*40，w 维度是 100，因此 T 维度为 512*100*40；拼接之后使用 CMVN 和 LDA 进行特征处理，根据特征和 UBM 获取每个话者的超向量，根据超向量 s、UBM、

F 模型，得到 i-vector 特征（S.Xue, 2014; M. Karafiat, 2011; N. Dehak, 2010）。最终得到 100 维的 i-Vector 特征，和 49 维的 MFCC 特征共同训练发音属性检测器组。

3.3 优化训练数据不平衡问题

建模时，汉语声母和韵母的建模分离和属性分类差异导致训练数据分布不平衡。例如，声母属性分类器中无用的标签“vowels”包含了近一半的训练数据。我们采用基于音素背景建模（phone-basedbackgroundmodel，PBM）的方法来解决这一问题，其关键是将无用分类和数据量庞大的分类进行多标签表示，就像在说话者或话语验证的方法，通过非属性类划分获得多标签。下图为本研究在属性检测器中使用 PBM 方法建模的示例图，该示例图为非属性类“vowels”化子标签的做法，以建模发音方式（PA）为例，横坐标为属性标签名，纵坐标为属性标签数量，蓝色是原始标签数据量，橘色是 PBM 算法进行数据平衡后的各标签数量。可以看到，蓝色部分“vowels”标签数量远大于其他标签，但是该标签在 PA 中没有任何意义，这样的数据分布会导致模型训练不平衡；而使用 PBM 后的橘色部分，将原标签“vowels”的数量平均分为四个子标签“vowels”、“vowels-a”、“vowels-b”、“vowels-c”，这样数据分布相对平滑。

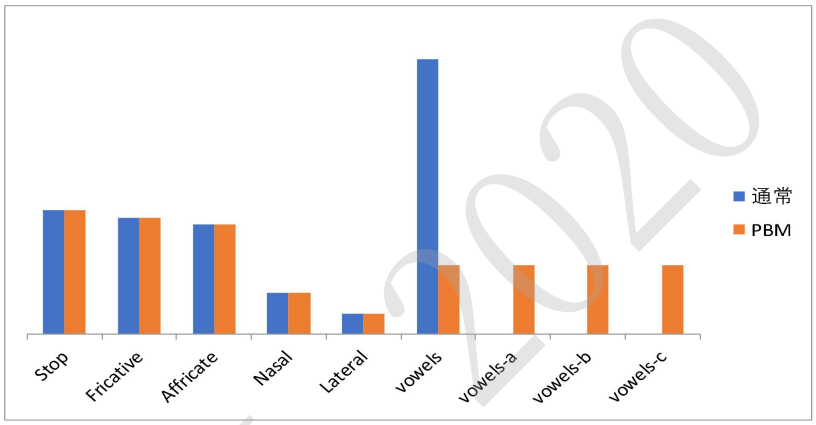


图 2.使用 PBM 方法对 PA 建模的数据分布

3.4 基于 FSA 的双语言属性检测器

众所周知，发音属性具备跨语言特性，为了探究基于 FSA 方法的跨语言属性识别能力，我们通过设计实验对照组，观察单语言和双语言训练的发音属性检测器在双语言属性识别任务中的性能对比。因为整合所有语言的发音属性本身比较难，我们还探索了单语言训练的属性检测器是否具备良好的多语言属性检测能力。但是 TDNN 和基于属性 HMM，两者同时建模发音属性的方法，有可能使得模型过于适应汉语发音习惯，而弱化发音属性原本的语言通用性质。因此，我们通过减少建模时使用的上下文的信息弱化模型对单个语言的适应性和依赖性，之后对比双语言训练的属性检测器的性能来验证这样做的可行性。弱化上下文信息的属性建模，我们采用 Monophone-HMM 和普通 DNN 模型作为对照组。

3.5 基于 FSA 的不同母语背景发音人的发音偏误检测

利用上述已被验证的语言之间共享发音属性的结论，可在发音偏误检测任务中用于建模发音偏误。由于二语者受到母语负迁移的影响，其发音偏误的发音属性常常会倾向于母语中的相似发音的发音属性，也就是说偏误发音实际上是介于二语者的母语和第二语言之间的发音。利用这一点，结合整合语言的属性检测器，可用于直接建模该发音人的发音偏误。理论上，在跨语言属性检测任务中性能良好的属性检测器，拥有描述不同母语背景学习者的发音偏误的能力。

为此，针对上述基于 FSA 的单语言和双语言训练的属性检测器，我们在不同母语背景学习者的发音偏误检测任务上进行测试，通过分析两组属性检测器在该任务上的性能，来验证是否

跨语言属性识别性能良好的属性检测器，也会拥有更好的描述发音偏误的能力。我们使用的两种二语语料测试集，分别为母语俄语的发音人和母语日语的发音人。

4 实验设计和结果

4.1 实验对照组

我们通过对比上下文相关的 HMM（triphone）组合 TDNN、上下文无关的 HMM（monophone）组合 TDNN、上下文无关的 HMM（monophone）组合 DNN 的三种建模发音属性的方法设置对照实验，在英语、汉语属性识别任务中的观察三个对照实验的性能，来测试单一语言训练数据下的三种方法建模发音属性时的跨语言能力。

同时，为了更直观地观察上述三个对照实验的效果，我们单独设置了一个对照实验，采用上下文相关的 HMM（triphone）组合 TDNN 的建模方法，数据上使用汉语和英语双语语料作为训练集，两语言训练数据量比例为 1:1，训练数据总量同上述三种方法一致，同样在英语、汉语属性识别任务中观察性能。

在发音偏误检测任务上，我们在两个测试集上设置了总计四组对照实验，两个测试集分别母语为俄语的学习者的中文语料、母语为日语的学习者的中文语料。四组对照组实验为：单语言训练属性检测器组+俄语背景学习者测试集；单语言训练属性检测器组+日语背景学习者测试集；双语言训练属性检测器组+俄语背景学习者测试集；双语言训练属性检测器组+日语背景学习者测试集。

最后一项对照组实验在于基于 FSA 的属性建模和基线属性建模两种方法的对比，我们针对二语学习任务设计了细颗粒度的发音属性定义并根据该定义建模了七种属性检测器，组成了前端属性检测器组。其中，只有两种属性与基线属性定义差距较大，即舌位前后、舌位高低，因此我们设置了两个对照组，分别观察在这两种属性上基于 FSA 和基线属性建模的两项后端任务的性能，即属性识别性能、偏误检测性能。下图为对照实验设计示意图。

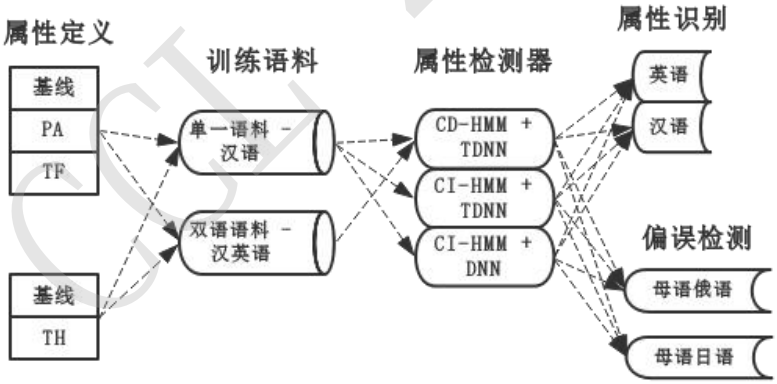


图 3.对照实验示意图

4.2 建模框架

本研究通过借鉴 ASAT 框架的整合思路，设计了基于 FSA 的建模框架。前端特征提取模块包含了一组属性分类器，用于提取属性后验概率，用于后端发音偏误检测任务，可以在不同维度上定义偏误检测，即超音段层级（如时长），音段层级（如音素替换），和次音段层级（如清化音素）（K.N.Stevens，2000）（G.Fant，1973）。本研究主要在次音段层级完成偏误检测实验，以及前端属性提取器的性能测试，具体过程框架如图 4 所示。

使用 MFCC 作为输入特征，设置对照组分别为 CD-HMMs、CI-HMMs，每个对照组包含七个基于发音属性的 HMM 模型；使用 MFCC 和 i-Vector 作为输入特征，两组基于属性的 HMM 做神经网络初始化，经过 PBM 的数据平衡后，建模基于属性的 TDNN 和 DNN，总计四个对照实验，每个对照实验七个模型；在每个前端分类器模型中，生成当前帧在该分类器中每个属性

的概率，即帧层级属性后验概率，作为前端输出。总计两个后端任务，将每个属性分类的帧层级后验概率用于评估基于 FSA 建模方法在中文和英文测试集上性能，之后进入强制对齐处理后转化为音素级后属性验概率进行次音段级发音错误检测，即中英文属性测试和次音段、音段偏误检测两项任务。

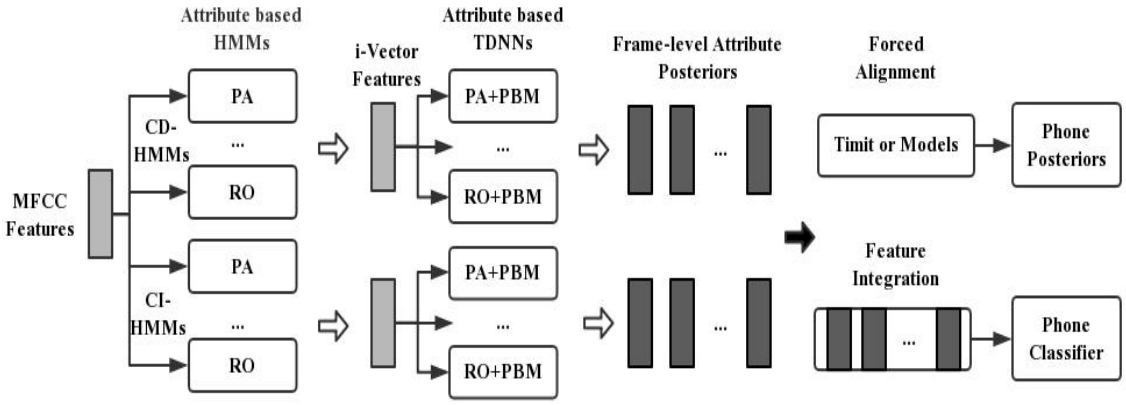


图 4.基于 FSA 的建模框架

根据后端任务的不同，分别对属性后验概率进行两种处理方式：在属性识别任务中，中文测试数据使用模型训练的强制对齐得到音素后验概率（Phone Posteriors），英文测试数据使用数据库自带的音素边界标注做对齐得到英文音素后验概率（Phone Posteriors）；在次音段偏误检测任务中，我们采用语音识别的整合过程（Phone Classifier）。此外，通过融合七种属性次音段级的偏误检测，我们完成了对不同母语背景的二语学习者语料库的音段级发音偏误检测，其中音频的音素边界信息是通过单独对二语语料本身建模后，经过强制对齐得到的。

4.3 发音属性检测

我们使用的语料库来自中国国家高新技术项目 863（S.Gao，2000），以及开源的 Aishell178 小时普通话语料库，英语语料库分别使用来自 LibriSpeech 和 Timit。单语言训练的属性检测器的训练数据共使用了 1800 名说话者（约 300 小时）的 25 万个话语进行声学建模，双语言训练的属性监测器的训练数据使用了 LibriSpeech 和 Aishel 两个语料库的数据，共 20000 条数据，约 300 小时，与单语言对照组的训练数据量保持一致，英语语料和汉语语料的比例为 1:1，充足的数据保证了基于 FSA 方法建模的鲁棒性。属性识别实验的的测试集有两个，一个是来自 Aishell 语料库的 6000 条中文数据，另一个是来自 Timit 的 6000 条英文数据。

我们对单语言训练的属性检测器在母语（Ch）和跨语言（En）发音属性检测任务上进行了评估；通过两种建模方法（上下文相关 CD、上下文无关 CI）和两个神经网络 DNN 模型、TDNN 模型，每个对照组包含三组对照实验（Triphone、Monophone、CI）。所有属性识别的实验结果如图 5 和图 6 所示。

由图 5 可知，上方三条曲线表示在汉语上测试（Ch）上表现出可靠的性能，即三个对照实验性能均在 80%以上，且上下文相关和 TDNN 组合建模（Triphone-Ch）的准确率，高于上下文无关和 DNN 组合（CI-Ch）建模的准确率。下面三条曲线在跨语言测试集（En）中表现出相对较低的检测准确率，尤其是元音部分，这表现出英语元音的结构与汉语差别很大，但是上下文无关和 DNN 组合（CI-En）建模的准确率，趋势上高于上下文相关和 TDNN 组合（Triphone-En）的准确率。经过更深入的观察，在跨语言属性检测任务中的多个属性检测器，如擦音（Fricative）和浊音（Voiced），可以获得较好的准确性（最高 93%和 78%）。我们还发现，在英语测试集上分类更精细的 TF 的属性集（见表 2）精度略优于 PA 分类（见表 1）。此外，由依赖于上下文的建模方法并不比上下文独立的建模优异，甚至 CI 方法在某些属性上也具有更高的效果，验

证了发音属性的语言独立性。

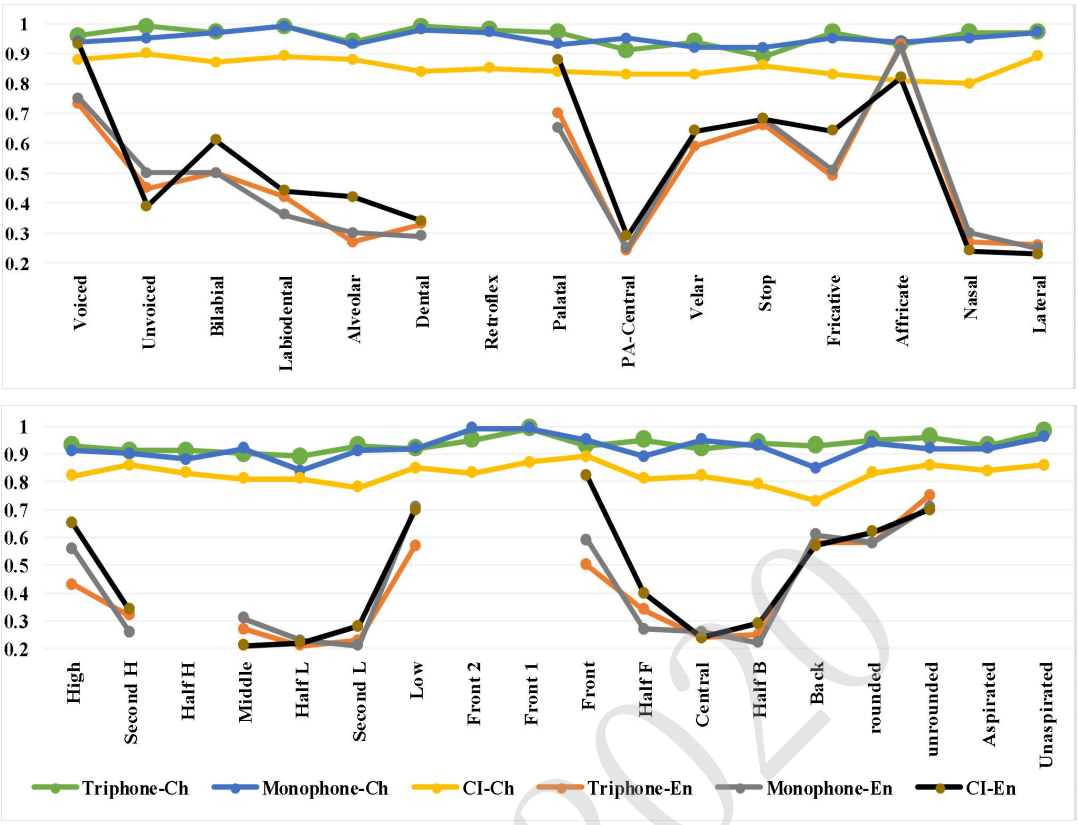
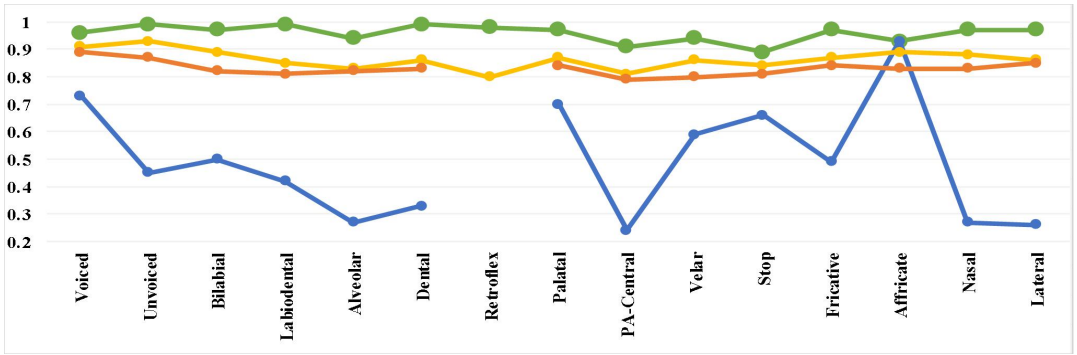


图 5.在汉语和英语上的基于 FSA 方法的检测准确率

我们同样对单语言训练的和双语言训练的属性检测器，在母语（Ch）和跨语言（En）发音属性检测任务上进行了评估，对应两个对照组（Ch、En），每个对照组包含两个对照实验（Monolingual、Bilingual），其中两个对照组上下文相关组合 TDNN，两个对照组除训练数据不同以外无其他差别。另外，由于英语中并没有 AS 属性，所以我们使用 PBM 方法平衡了双语训练集数据来训练 AS 属性检测器。所有属性识别的实验结果如图 6 所示。如图，准确率最高的两个曲线为汉语测试集上的属性识别结果（Ch），识别准确率在 80%以上，且单语言训练的属性检测器识别准确率（Monolingual-Ch）均高于双语言训练的属性检测器识别准确率（Bilingual-Ch）。图中下两条曲线反映了英语属性识别对照组的情况，其中双语言属性检测器识别准确率（Bilingual-En）远高于单语言属性检测器识别准确率（Monolingual-En）。



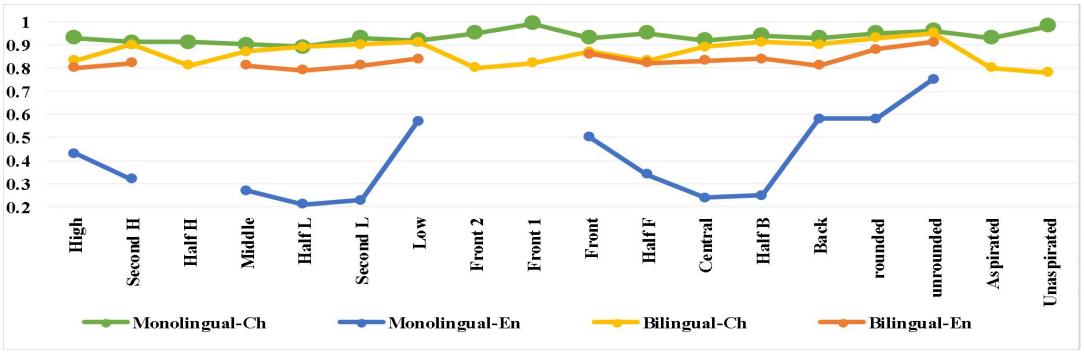


图 6.在汉语和英语上的基于 FSA 方法的检测准确率

4.4 发音偏误检测

发音偏误检测任务采用的二语语音数据库，使用北京语言大学汉语言语料库（J.-S.Zhang，2010），其中包含母语为俄语的普通话学习者的 1000 条语音，和母语为日语的普通话学习者 1000 条语言。为了在次音段级和音段级上检测发音偏误，我们使用了两个指标，即 F-score 和诊断准确率（DA）来评估发音错误检测的性能。

$$DA = \frac{N_M + N_C}{N} * 100\ \% \qquad \text{公式（4.1）}$$

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad \text{公式（4.2）}$$

其中 N_M 为检测到的真实偏误数，检测结果与人工标注一致。N_C 是系统检测到的真实正确发音的个数。Precision 真实偏误数和所有检测到的发音错误的数量的比值，称为准确率。其中 Recall 为真实偏误数和测试集中发音错误总数的比值，称为召回率。N 为测试集中音素或属性的个数。

在单语言训练的属性检测器组的对照实验中，我们选取了 7 个具有最好属性识别性能的分类器，并在两种测试集上对次音段级偏误检测性能进行评估，之后将它们整合至音段级的偏误检测中。我们可以看到，不论是母语为俄语还是日语的发音人测试集，基于 FSA 的方法都可以很好地检测出不同的发音属性的偏误，尤其单语言训练的属性检测器组（Monolingual-DA），次音段级诊断准确率均在 83%以上。更精细地体现汉语舌位变化的 TF、PA、TH，不论是在两个测试上（母语为俄语或日语）还是单语言或双语言训练（Monolingual-DA，Bilingual-DA）的属性检测器，表现均高于基线的 T-T 和 H-H。

	VO	AS	MA	PA	TH	TF	RO	T-T	H-H
Monolingual-DA	89.4%	89.0%	87.2%	83.4%	85.9%	84.3%	88.2%	82.7%	85.2%
Bilingual-DA	83.6%	82.4%	78.8%	72.6%	74.4%	73.9%	82.9%	72.2%	73.6%

表 3.母语俄语学习者次音段偏误检测

	VO	AS	MA	PA	TH	TF	RO	T-T	H-H
Monolingual-DA	91.4%	90.2%	88.6%	85.2%	86.7%	87.0%	90.7%	83.9%	85.7%
Bilingual-DA	84.7%	83.4%	81.2%	74.1%	76.0%	75.0%	83.6%	73.6%	74.9%

表 4.母语日语学习者次音段偏误检测

在母语为俄语的二语者的测试集上，通过对比，单语言训练和双语言训练的属性检测器组的整体诊断准确率（Monolingual-DA，Bilingual-DA），我们发现双语言训练的属性检测器组的偏误检测诊断准确率低于单语言训练的属性检测器。在母语为日语的二语者的测试集上的偏误检测结果，总体上母语为日语的学习者的次音段偏误检测准确率比母语为俄语的学习者要高， 可能是因为母语为日语的学习者的汉语总体水平高于母语为俄语的学习者。通过对比关于单语言训练和双语言训练的属性检测器组的诊断准确率（Monolingual-DA，Bilingual-DA），同样地， 双语言训练的属性检测器总体表现低于单语言属性检测器。

	FSA-based（M/B）		Segment-based（M）	FSA-based（M/B）		Segment-based（M）
F-score	71.5%	61.2%	63.5%	74.7%	62.8%	67.9%
DA	86.5%	78.4%	84.3%	88.5%	79.7%	85.8%

表 5.母语俄语/日语学习者音段偏误检测

将上述单语言和双语言训练的属性检测器组分别整合后，与同数据量训练（Aishell，约 300 小时）的基于音段的偏误检测相比（Monolingual，M），在两种母语背景学习者的测试集上，基于FSA 的偏误检测诊断准确率更高，F-score 更高，验证了本研究提出方法的有效性。此外，在母语为日语的发音人测试集中，音段偏误检测的性能均优于在母语为俄语的发音人测试集中的性能，包括基线系统（Segment-based）的基于音段的偏误检测诊断准确率；双语言属性检测器组整合后（Bilingual，B），用于音段偏误检测，在 DA 和 F-score 上低于单语言属性检测器组整合后的结果，这与次音段偏误检测中的结果一致。

5 结论

我们提出了一种基于细颗粒度发音属性（FSA） 识别并在发音偏误检测中应用。 实验结果表明， 在使用单一语言训练时，该方法提取了可靠的帧层级发音属性的准确率，均在 90%以上；在跨语言测试中，通过修改建模时使用的上下文信息降低了检测器在汉语上的过度适应， 建模时使用的上下文信息越少，单语言属性检测器性能越好，验证了发音属性的跨语言特性； 但是，使用上下文信息最少的属性检测器组，跨语言测试的准确率也远低于双语言属性检测器在英语属性识别任务中的性能，证明语言间音素结构的巨大差异依然有很大影响。在汉语属性识别任务中，单语言训练相比双语言训练的属性检测器组，准确率平均高出 7%，这表明双语言属性检测器，没有很好地表现出发音属性的语言独立性。相比单语言训练，双语言属性检测器组在英语属性识别任务中的性能提升明显，体现了属性的语言通用性。

在二语学习者的偏误检测实验中，使用基于FSA 的方法相比于传统发音属性定义的基线系统，次音段级别偏误检测任务中都表现了更优的性能，表明基于FSA 的方法在偏误检测任务中更能体现汉语语言发音的特点；同时，同数据量训练的基于发音属性的方法（单语言）比起基于音段的方法，在音段偏误检测任务中获得了更好的检测性能，近一步验证了基于 FSA 方法的有效性。

理论上，该方法可以应用于任何母语背景的学习者，我们通过在母语背景为俄语、日语的发音人语料库上的发音偏误检测，测试双语言训练相比单语言训练的的属性检测器，是否能拥有更好的描述发音偏误的能力，实验结果显示，单语言训练的属性检测器性能更优。经过分析， 可能由于双语训练使用的第二语料库，并非使用发音人的第一语言，即日语和俄语；双语属性检测器在汉语属性识别任务中准确率低于单语言属性检测器，即没有体现属性的语言独立性。

致谢

本论文受到国家社科基金项目（18BYY124），语言资源高精尖创新中心项目 (KYR17005)，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）（19PT04），北京语言大学一流学科团队支持计划（GF201906）项目资助。本文通讯作者为解焱陆。

参考文献

Ambra Neri, Catia Cucchiarini, Helmer Strik. ASR-based corrective feedback on pronunciation: does it really work[J]. proceedings of interspeech icslp pittsburgh pa september, 2006:1982-1985.

C. Zhang, Y. Liu, and C.-H. Lee, “Detection-based accented speech recognition using articulatory features,” in Proc. ASRU, 2011.

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[C]// INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010. DBLP, 2010.

Duan R , Kawahara T , Dantsuji M , et al. Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data[C]// ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017. Florian Metze, Articulatory Features for Conversational Speech Recognition, Ph.D. thesis, Karlsruhe, Univ., Diss., 2005, 2005

G. Fant, Speech Sounds and Features. Cambridge, MA, MIT Press, 1973.

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]//SLaTE. 2009: 45-48.

J.-S. Zhang, W. Li, et al, “A Study On Functional Loads of Phonetic Contrasts Under Context Based On Mutual Information of Chinese Text And Phonemes,” in Proc. ISCSLP, 2010.

K. N. Stevens, Acoustic Phonetics. Cambridge, MA, MIT Press, 2000.

Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. H. 2016. Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. IEEE International Conference on Acoustics, Speech and Signal Processing (pp.6135-6139). IEEE.

M. Karafiat, L. Burget, P. Matejka, O. Glembek, and J. Cernocky, “iVector-based discriminative adaptation for automatic speech recognition,” in 2011 IEEE Workshop on Automatic Speech Recognition & Understanding.

Muller M , Franke J , Waibel A , et al. Towards phoneme inventory discovery for documentation of unwritten languages[C]// ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, Front-end factor analysis for speaker verification IEEE Transactions on Audio Speech & Language Processing, vol. 19, no.4, pp. 788-798, 2011IEEE, Dec. 2011, pp.

S. Gao, et al, “Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR,” in Proc. ICSLP, 2000.

S. M. Siniscalchi et al, “Toward a detector-based universal phone recognizer,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, Nevada, USA, Mar./Apr. 2008, pp. 4261–426

S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q.-F. Liu, “Fast Adaptation of Deep Neural Network based on Discriminant Codes for Speech Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, no. 99, pp.

Waibel A , Hanazawa T , Hinton G E , et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics Speech and Signal Processing, 1989, 37(3):328-339.