北京大学学报(自然科学版) 第 54 卷 第 2 期 2018 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 2 (Mar. 2018)
doi: 10.13209/j.0479-8023.2017.152

基于发音特征的汉语发音偏误自动标注

魏星 王玮 陈静萍 解焱陆† 张劲松

语言资源高精尖创新中心，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 针对发音偏误检测系统语音标注费时、费力和标注不一致的问题，基于发音特征，构建偏误检测系统，给出Top-N的识别结果，通过praat软件呈现机器初步标注文本，在此基础上进行人工二次标注。实验结果表明，与单纯的人工标注相比，所提出的自动标注加人工二次标注方法在标注一致性上从80.7%提高到92.48%，平均每个句子的标注时间从10分钟减少到3分钟。所提方法有效地提高了人工标注的效率，可以在有限时间内为识别系统提供更多可靠的标注语料。

关键词 发音特征；发音偏误趋势；自动标注

中图分类号 TP391

A Study of Articulatory Features Based Detection of Mandrain Pronunciation Erroneous Tendency for Automatic Annotation

WEI Xing, WANG Wei, CHEN Jingping, XIE Yanlu†, ZHANG Jinsong

Advanced Innovation Center for Language Resource and Intelligence Research Funds of State Language Commission, School of Information Science, Beijing Language and Culture University, Beijing 100083; † Corresponding author, E-mail: xieyanlu@blcu.edu.cn

Abstract For the purpose of relieving the time cost and inconformity in annotation, the authors use an articulatory features based mispronunciation detection system to give an Top-N feedback and use this feedback to assist manual annotation. As a result, the consistency rate of phoneme labels in proposed system increases from 80.7% to 92.48%. In addition, the time cost for annotating each sentence reduce from 10 to 3 minutes. The results indicate that proposed automatic annotation system is practical, and there is also a room for further improvement.

Key words articulatory features (AFs); pronunciation erroneous tendency (PET); automatic annotation

近年来，随着机器学习和计算机技术的发展，自动语音识别(ASR)技术成为当前研究热点之一。有标注的语料库在语音合成、语音识别、语音分析等领域发挥着日益重要的作用。为大规模语音语料库添加标注是一项需要投入大量人力资源的任务，长时间的连续工作不可避免地造成标注人的疲劳和倦怠，同时标注人所接受的语音学专业训练水平、对语音学知识的把握以及生理、心理因素的共同影响，都会造成主观误差，影响标注结果[1]。因此，必须发展语音自动标注系统。

语音语料库的标注方法一般有自动标注和人工标注两种，或两者相结合的方法，例如先用ASR系统对语音数据进行自动标注，然后再进行人工校正[2]。朱维彬等[1]认为，语音自动标注系统有两条技术路线：1) 基于统计模型，基础是样本量足够大的附手工标注信息的语料库；2) 基于语言学模型，出发点是由语言声学知识总结的先验性规则。

由于自动标注的准确性不如人工标注，现有的ASR 系统无法实现语音语料库的全自动标注，标注工作往往通过自动标注和人工标注相结合的方式完成。对未标注的语料库，一般先用自动标注的方法标注音素层信息，再由专业标注人员进行校对和

国家语言文字工作委员会科研项目(ZDI135-51)、北京语言大学梧桐创新平台项目(16PT05)、北京语言大学研究生创新基金项目(17YCX140)和语言资源高精尖创新中心项目(451122500)资助

收稿日期: 2017–06–05; 修回日期: 2017–09–05; 网络出版日期: 2017–11–04

标注[3]。

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和主要发音单元之间的对应关系[4]。在霍普斯金大学 2006 年暑期语音研讨会上，国外主流语音实验室就如何将发音特征引入语音识别系统进行了探讨，结果表明，将发音特征引入语音识别系统有助于系统识别性能的改善[5−6]。与语音的频谱或倒谱特征相比，发音特征能够更加直观地反映发音器官的变化规律。首先，发音特征使音素间的协同发音现象能更自然地被建模，为分析协同发音以及对音素序列的恢复提供更多的潜在信息。其次，发音特征独立于声学环境的变化，不易受噪声之类的声学环境的影响[7]。与常规的声韵母单元相比，采用发音特征建模可以更好地描述发音偏误类型，有助于提升发音偏误检测系统的性能。发音特征在语音识别中的这些优势已受到越来越多的关注[5−6,8]。

本文从发音特征的角度对发音偏误建模，通过偏误自动检测方法，对面向二语学习者的中介语语料库自动标注，在此基础上进行人工校对和标注。

1 发音偏误自动检测

1.1 发音特征归类

对于二语学习者来说，发音偏误中有一些非此即彼的音位替换，但更多的是似 A 似 B 式的音素不准[9]。Cao 等[10]根据发音位置和发音方式等的不准确性，定义了相应的发音偏误趋势，包括高化、低化、前化和后化等 64 种。Duan 等[11]以及 Gao 等[12]的研究表明，将发音偏误趋势加入检测中，不仅可以检测出学习者的偏误发音，也能向二语学习者给出发音位置和发音方式等的反馈信息。

本文要标注的是发音偏误，因此将音素按照发音方式、发音位置、是否送气和清浊音分为 4 类，发音特征与音素的对应关系如表 1 所示[13]。

表 1 发音特征与音素对应关系

Table 1 Articulatory features and their associated phones

类别	发音特征	音素
发音位置	双唇音	b, p, m
	唇齿音	f
	齿龈音	d, t, l, n
	齿音	c, s, z, ii
	卷舌音	zh, ch, sh, r, er, iii
	腭音	j, q, x, a, o, e, i, u, v
	软腭音	g, k, h, ng
发音方式	塞音	b, p, d, t, g, k
	擦音	f, s, sh, r, x, h
	塞擦音	z, zh, c, ch, j, q
	鼻音	m, n, ng
	边音	l
	N/A	a, o, e, I, ii, iii, u, v, er
是否送气	送气音	p, t, k, c, ch, q
	不送气音	b, d, g, z, zh, j
	N/A	f, h, l, m, n, r, s, sh, x, ng, a, o, e, I, ii, iii, u, v, er
清浊音	浊音	m, n, l, r, ng, a, o, e, I, ii, iii, u, v, er
	清音	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s
静音	静音	sil

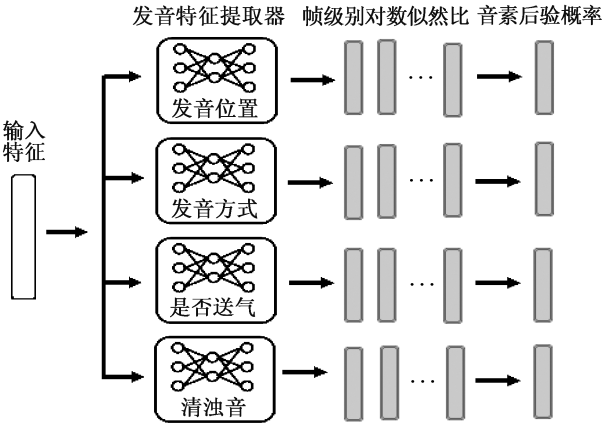

图 1 发音特征提取流程

Fig. 1 Flow chart of articulatory features extractors

1.2 发音偏误检测框架

我们使用基于统计语音识别的检测框架来实现发音偏误的自动检测功能，整个检测框架如图 1 所示。首先将提取的语音帧分别输入每个发音特征提取器中，然后从发音特征提取器中输出每个 senone 的似然值，并根据式(1)计算每个音素的发音特征后验概率，之后根据后验概率大小排序，得到 Top-N 的检测结果，最后将 Top-N 的检测结果生成标注文本，用于标注。

本实验计算后验概率的方法基于文献[14]，每一个音素都用式(1)计算后验概率:

$$\log P(p|\mathrm{O};t_\mathrm{s},t_\mathrm{e})=\frac{1}{t_\mathrm{e}-t_\mathrm{s}}\sum_{t_\mathrm{s}}^{t_\mathrm{e}}\log\sum_{s\in p}P(s|O_t),\quad(1)$$

Q_t是在 t 时刻的输入特征；t_s 和 t_e是音素 p 的起始和终止时间，通过强制对齐得到；$P(s|O_t)$是帧级别的对数似然值；$\{s\in p\}$是所有属于音素 p 的帧的集合。

2　标注

2.1　标注规范

在进行偏误标注时，首先应对目标学习者可能出现的发音偏误有较全面的了解。以日本学生为例，常见的偏误有是否送气的混淆、前后鼻音的混淆、r 与 l 混淆、sh 与 x 混淆等，日本学生汉语中介语语音语料库的标注符号至少要覆盖所有这些偏误类型[15−17]。此外，非 A 即 B 以及似 A 似 B 一类的差异在标注系统里也必须考虑到[18]。

一般的语音语料库的标注都用国际音标(IPA)做语音学标注，未对发音偏误做任何标注，这使得语料库的实用性大大受限，尤其是针对 CAPT 系统[19]。本实验采用的标注方案借鉴 Cao 等[18]提出的中介语语料库标注方案，部分规范如表 2 所示。这套标注方案让使用者通过标注文本就能获得学习者的偏误类型。

2.2　标注方法

本实验有两位语音学研究生参与人工校对与标注，流程如图 2 所示。正式标注前，标注人会拿到一份详细的标注规范，并挑选几个句子进行试标注。图 3 中第 4~7 层分别为 4 个特征提取器的检测结果，为方便标注，仅给出 Top-N 中不一致的部分，即判断为偏误的发音特征标签。由于采用独立训练的方法，4 个模型的音素边界均通过强制对齐获得，导致边界没有完全对齐，为方便查找，按照从前往后的顺序标记发音特征标签的序号，如图 3 中 4~7 层的 1~10。标注人只需校对后 4 层给出标签部分，将偏误符号标注在第 3 层对应的“{ }”中，如果标注人判断提示处无偏误，则忽略提示。如图 3 中第 3 层的“t{;}”，在第 6 层给出提示 uas (即 unaspirated 的缩写)，表示“t”被检测为一个不送气音，存在送气不足，所以在“t{ }”内标上短化符号“;”，后面的“ian{ }”虽然给出提示，但标注人判断此处无偏误，则忽略提示。标注过程不限制单句话标注时间，但会统计总时间作为参考。所有的标注工作都使用“Praat 6.0.26”完成[20]。

表 2　汉语中介语语料库音段标注规范(BLCU-CAPL-1)
Table 2　Inter-Chinese corpus annotation standard

类型	标注符号	偏误举例	备注/说明
前化	+	e{+}n	e 的舌位靠前
后化	–	n{–}	前鼻音发音近似后鼻音
短化	;	p{;}	p 送气时长不够
圆唇化	o	e{o}	e 被发成了圆唇音
展唇化	w	u{w}	u 被发成了不圆唇音
舌叶化	sh	sh{sh}	sh 被发成了 x
边音化	l	r{l}	r 被发成了 l

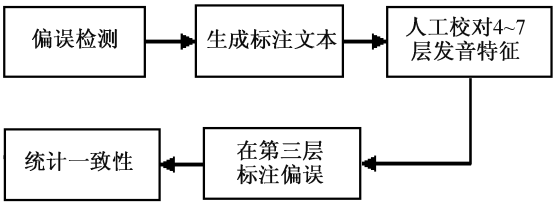

图 2　标注流程
Fig. 2　Schematic diagram of annotation

3　实验及结果分析

3.1　实验语料

实验所用语料库来自北京语言大学中介语语料库。我们选取 7 位日本女学生的连续语音，每人约 301 句话(日常用语)。由 6 位语音学专业的研究生对其进行发音偏误的交叉标注，当出现不一致时，请语音学专家进行判定。语料统计结果见表 3，其中约 80%的数据用于训练，其余用于测试。

本实验使用有监督的训练方法，基于深度神经网络(DNN)分别建立 4 个发音特征提取器，声学特征使用 13 维的 MFCC 特征以及其一阶、二阶差分，以 20 ms 为窗长，10 ms 为帧移提取。DNN-HMM 模型的输入是当前帧以及前后各 5 帧，共 11 帧构成的特征向量。由于在汉语普通话中，与韵母相比，声母更容易导致发音偏误，所以本文主要针对声母的偏误。

3.2　评价指标

3.2.1　偏误检测系统评价指标

实验结果有 4 种：1) 正确接受(TA)，正确发音检测为正确发音的个数；2) 正确拒绝(TR)，偏误发音检测为偏误发音的个数；3) 错误接受(FA)，偏误发音检测为正确发音的个数；4) 错误拒绝(FR)，正确发音检测为偏误发音的个数。

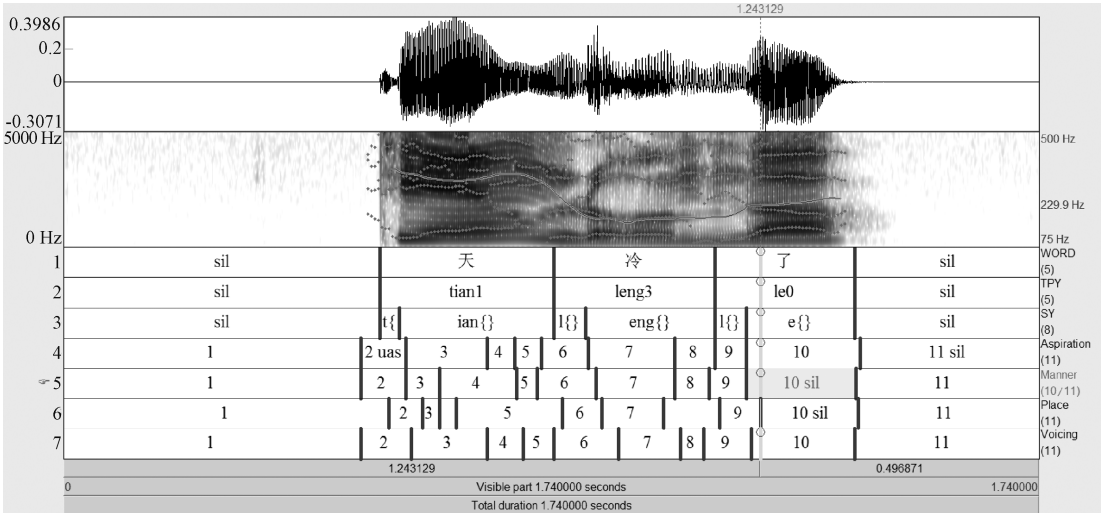

图 3 标注示例

Fig. 3 A real annotation example

表 3 日语中介语语料统计

Table 3 Japanese L2 inter-Chinese corpus

文本	录音人	句子总数	音素总数	每句话平均音素数	标注者人数	每句话标注者人数
301 句日常用语	7 个日本女学生	1899	26431	14	6	2

根据这 4 种检测结果，可以计算 3 种常见的评价指标。

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，

$$\mathrm{FAR}=\frac{\mathrm{FA}}{\mathrm{FA}+\mathrm{TR}}\times 100\%\text{。} \tag{2}$$

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，

$$\mathrm{FRR}=\frac{\mathrm{FR}}{\mathrm{FR}+\mathrm{TA}}\times 100\%\text{。} \tag{3}$$

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，

$$\mathrm{DA}=\frac{\mathrm{TA}+\mathrm{TR}}{\mathrm{TA}+\mathrm{TR}+\mathrm{FA}+\mathrm{FR}}\times 100\%\text{。} \tag{4}$$

3.2.2 标注评价指标

实验中有两位语音学研究生在偏误检测结果基础上进行人工校对，主要评价指标为标注人之间的一致性，根据标注内容，可以分为以下 4 种。1) 一致正确(CC): 两位标注人均认为发音正确。2) 一致偏误(CM): 两位标注人所标偏误符号一致。3) 不一致偏误(IM): 两位标注人所标偏误符号不一致。4) 争议偏误(WM): 两位标注人中仅有一人判断为偏误。

3.3 实验结果

3.3.1 偏误检测结果

我们采用有监督的方式训练了 4 个基于 DNN-HMM 的发音特征提取器，并采用上述评价指标来衡量系统性能。从计算机辅助发音训练(CAPT)的目的出发，应避免将学习者的正确发音判断成偏误发音，否则会打击学习者的积极性。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，检测结果见表 4。

通过对实验结果的进一步分析，我们发现系统整体检测效果尚可，但对舌叶化、闪拍化和卷舌化的检测效果不太明显，对其他主要类型的发音偏误

表 4 偏误检测结果

Table 4 Mispronunciation detection result

发音特征模型	FRR/%	FAR/%	DA/%
发音位置	7.5	39.7	84.5
发音方式	6.5	36.3	86.3
是否送气	6.9	37.4	85.7
清浊音	6.7	36.5	85.9

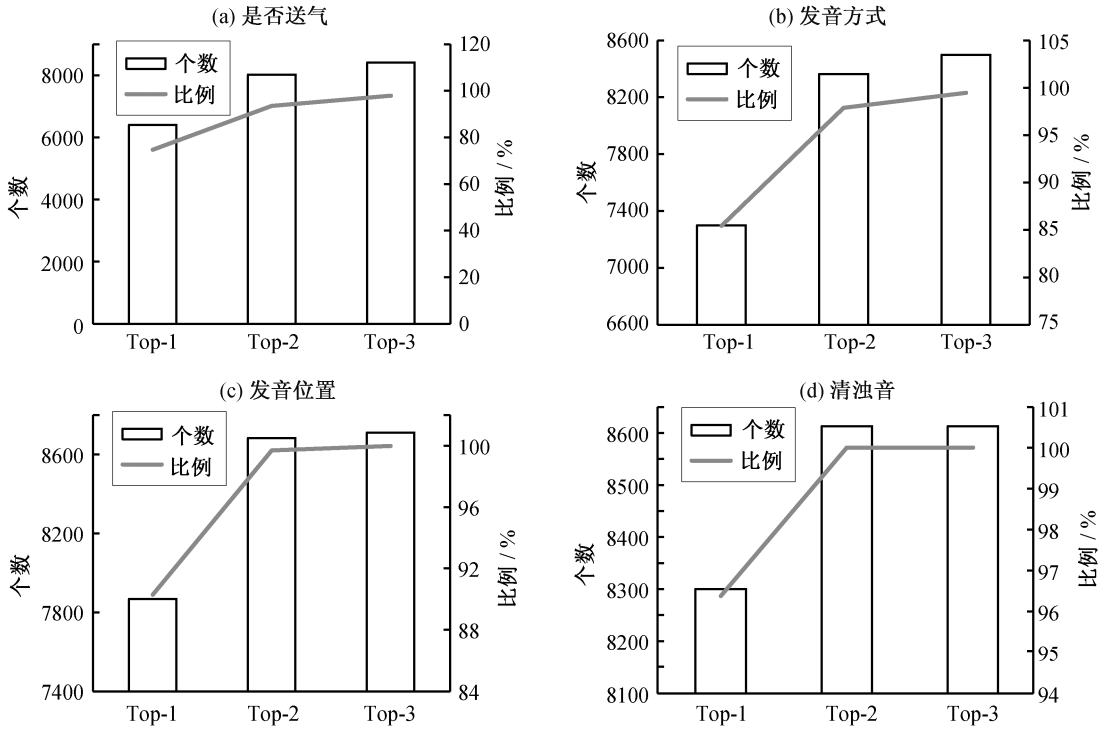

图 4　Top-N 结果
Fig. 4　Top-N result

检测效果比较好。实验还给出基于特征后验概率的 Top-N 的排序结果，反映发音特征检测结果与原始文本的一致性，如图 4 所示。可以发现，4 个模型 Top-1的一致性最好的是清浊音(96.37%)，最差的是发音位置(74.58%)，4个模型 Top-2 的一致性均达到 93%以上，说明 Top-N 的结果具有较好的参考性，对标注人标注偏误有一定的辅助作用。对比 Top-1 结果与原始标注可知，二者一致性达到约 82%，其中约 75%为一致正确，剩下 7%为一致偏误。

3.3.2　标注结果

经过统计分析，本实验中两位标注人的一致性达到 92.48%，比原人工标注 80.7%的一致性有较大提高。除对比两位标注人之间的一致性外，还与之前的标注结果进行一致性对比，结果如图 5 所示。

图 5 中 jp 和 ww 为本实验两位标注人的结果，ori 为之前的人工标注一致性结果。从图 5 可以发现，前人的一致性为 80.7%，本实验中两位标注人的一致性达到 92.48%，与前人有较大的提升。同时，本实验标注与前人标注的一致性也分别达到 79.81%和 79.1%，与原标注的一致性基本上相当。jp 与 ori 以及 ww 与 ori 的争议偏误比例分别达到 15.3%和 16.2%，说明两位标注人对自动偏误检测的结果判断不一致。这种情况主要由两个原因导致，一方面是标注人之间的个体差异，有些发音偏误不够清晰明确，虽然有偏误检测结果，但标注人在进行人工校对时可能听不出此处的偏误，因此认定为无偏误；另一方面与偏误检测系统性能相关，统计发现，争议偏误主要出现在舌叶化、卷舌化和闪拍化等几种偏误，可能是由于数据稀疏导致几种偏误类型的检测性能不够好。本实验中，两位标注人标注一句话平均需要 3 分钟，与原来的平均 10 分钟一句话相比，可以节约大量的时间成本，能

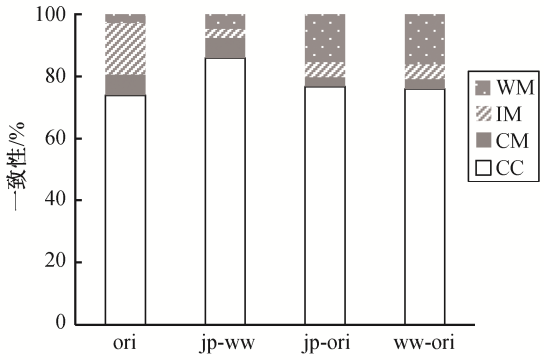

图 5　标注一致性结果
Fig. 5　Result of annotation consistency

在有限时间内为系统提供更多可靠的标注语料。

4 总结与展望

为缓解人工标注语料库时存在的费时、费力以及一致性不高的问题，本文引入基于发音特征的自动标注方法。首先训练 4 个基于 DNN-HMM 的发音特征提取器，然后利用输出的帧级别的似然值计算音素后验概率，再根据后验概率大小排序得到 Top-N 的分类结果，最后将结果反馈给标注人进行标注，并统计一致性。实验对比了前人标注结果，虽然争议偏误略微提高，但总体上一致性更高，达到 92.48%。此外，每句话的平均标注时间也从原来的 10 分钟降到 3 分钟，可以较大程度地缓解人工语料库标注费时、费力和一致性不高的问题。从实验结果来看，本实验中 4 个模型独立给出反馈结果对标注人造成一定程度的困惑，今后可以采用 ASAT 框架来解决这个问题。将来也需要加大训练数据规模，进一步改善声学模型，提高检测正确率。

参考文献

[1] 朱维彬，张家录. 汉语语音数据库的标注 // 全国人机语音通讯会议. 北京, 1996: 350–353

[2] 章森，华绍和. 普通话广播语音的多层次标注与检索. 中文信息学报, 2007, 21(4): 97–104

[3] Bonaventura P, Howarth P, Menzel W. Phonetic annotation of a non-native speech corpus // Proceedings of the Workshop on Integrating Speech Technology in (Language) Learning. Dundee, 2000: 10–17

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

[5] Livescu K, Cetin O, Hasegawa M, et al. Articulatory fature-based methods for acoustic and audio-visual speech recognition: JHU Summer Workshop Final Report // ICASSP. Honolulu, 2007: 621–624

[6] Cetin O, Cantor A, King S, et al. An articulatory fature-based tandem approach and factored tandem observation modeling // ICASSP. Honolulu, 2007: 645–648

[7] 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

[8] 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

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

[10] 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 // INTERSPEECH. Makuhari, 2010: 1922–1925

[11] Duan Richeng, Zhang Jinsong, Cao Wen, et al. A preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners // INTERSPEECH. Singapore, 2014: 1478–1481

[12] Gao Yingming, Xie Yanlu, Cao Wen, et al. A study on robust detection of pronunciation erroneous tendency based on deep neural network // INTERSPEECH. Dresden, 2015: 693–696

[13] Li W, Siniscalchi S M, Chen N F, et al. Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling // ICASSP. Shanghai, 2016: 6135–6139

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

[15] 朱川. 汉日语音对比试验研究. 语言教学与研究, 1981(2): 42–56

[16] 曹文. 汉语语音教程. 北京: 北京语言大学出版社, 2002

[17] 王蕴佳. 日本学习者感知和产生汉语普通话鼻音韵母的实验研究. 世界汉语教学, 2002(2): 47–60

[18] 曹文，张劲松. 面向计算机辅助正音的汉语中介语语料库的创制与标注. 语言文字应用, 2009(4): 122–131

[19] Hincks R. Speech recognition for language teaching and evaluation: a study of existing commercial products // Proceedings of ICSLP. Denver, 2002: 733–736

[20] Boersma P, Weenink D. Praat: doing phonetics by computer [EB/OL]. [2017–05–14]. http://www.fon.hum.uva.nl/praat