自动发音错误检测中基于最大化 $F1$ 值准则的区分性特征补偿训练算法

黄 浩[1],徐海华[2],王美慧[1],吾守尔·斯拉木[1]
(1.新疆大学信息科学与工程学院,新疆乌鲁木齐 830046;2.南洋理工大学 Temasek 实验室,新加坡 639798)

摘 要: 为提高自动发音错误检测性能,提出一种区分性特征补偿训练算法.该方法将高斯后验概率矢量经过线性变换后作为偏移量补偿至传统的谱特征.将经过正确度标注的语音数据库上的发音错误检测 $F1$ 值的最大化作为变换参数的训练准则.推导了目标函数对变换参数的偏导数公式,并利用无约束参数优化例程 L-BFGS 更新变换参数.发音错误检测实验表明该方法能够有效增大训练和测试集的 $F1$ 值.并且训练和测试集的精确度、召回率也都有明显提高.在特征优化的基础上进行模型参数训练,检错性能较单独的区分性特征训练、单独的区分性模型训练都有进一步改进.

关键词: 自动发音错误检测; $F1$ 值; 区分性训练; 特征; 计算机辅助语言学习
中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2015)07-1294-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.07.007

Maximum $F1$-Score Criterion Based Discriminative Feature Compensation Training Algorithm for Automatic Mispronunciation Detection

HUANG Hao[1], XU Hai-hua[2], WANG Xian-hui[1], Wushour Silamu[1]
(1. *School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;
2. *Temasek Laboratories, Nanyang Technological University, Singapore* 639798)

Abstract: To improve the performance of automatic mispronunciation detection, a discriminative feature compensation training algorithm is proposed. The method is to train a matrix projecting from posteriors of Gaussians to a normal size feature space, and then to add the projected features to traditional spectral features. The matrix is trained according to maximum $F1$-score criterion, which aims at maximizing the empirical mispronunciation detection $F1$-score on the annotated speech database. Mispronunciation detection experiments have shown the method is effective in increasing $F1$-score, precision and recall on both the training data and evaluation data. It is also shown model parameter discriminative training on new features obtained further improvements over both model training and feature training.

Key words: automatic mispronunciation detection; $F1$-score; discriminative training; feature; computer-assisted language learning

1 引言

计算机辅助语言学习是利用语音语言技术辅助学生学习语言的技术,自动发音错误检测是其中一个重要形式,其目的在于自动指出说话人的发音错误.近年来,各机构都展开了研究,产生了一系列方法[1~9].发音错误检测总体上基于语音识别技术,在声学模型的建立方面,尽管研究人员对不同的模型进行了尝试[1,7,8,9],高斯混合模型-隐马尔可夫模型(Hidden Markov Model, HMM)仍是常用建模方法,而 GOP(Goodness of Pronunciation)[2]则是描述发音定量得分的经典算法.

语音识别中,利用区分性准则优化模型参数已证明较最大似然(Maximum Likelihood, ML)准则在识别率上有明显的改进.这些区分性准则有最小分类错误[10]、最大互信息[11]以及最小音子错误[12].这些准则的目的在于减小训练集的期望误识率,进而减小测试集误识率.而发音错误检测中,系统应尽可能逼近专家的检错标准,所以发音检错中性能评价指标与识别任务明显不同,

收稿日期:2013-12-26;修回日期:2014-09-24;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61365005, No.60965002)

直接使用上述准则优化参数不一定能达到检错指标下的最佳性能.针对该问题,我们在文献[13]中提出了最大化 $F1$ 值准则(Maximum $F1$-Score Criterion,MFC)的区分性模型训练算法,模型优化的目的在于提高对经过正确度标注数据的检错 $F1$ 值.实验表明,该算法能够显著提高检错 $F1$ 值.

语音识别中,根据区分性准则优化特征已证明是改进识别性能的有效方法,常见算法有:最大互信息-立体声环境线性分段补偿[14],特征空间最小音子错误[15],基于最小音子错误的区域相关变换[16]等.根据上述思想,提出发音错误检测的区分性特征补偿训练算法.该方法利用线性变换将观察矢量的高斯后验概率映射为偏移量并与谱特征叠加,变换根据 MFC 准则优化.给出目标函数对变换的偏导数计算公式,并利用无约束优化方法 L-BFGS (Limited-memory-BFGS)[17]更新变换.发音检错表明该方法能够有效提高训练和测试数据检错的 $F1$ 值.同时训练和测试数据上的精确度、召回率都有明显改进.在优化后的特征上进行区分性模型训练,检错性能又得到进一步提升.

2　最大化 F1 值准则

发音错误检测的目的在于判断音素发音是否正确.$F1$ 值是评价系统性能的常见指标,通过机器和评测员的检错结果的精确度(Precision)和召回率(Recall)计算:

$$F1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{1}$$

精确度和召回率分别为

$$\text{Precision}=\frac{N_{\text{WW}}}{N_D},\ \text{Recall}=\frac{N_{\text{WW}}}{N_W} \tag{2}$$

N_{WW}为同时被人工和机器判为错误的音素数目,N_D 是机器检测为错误的数目,N_W 是人工标注为错误的数目.设检错数据有 R 条语句,每条语句的特征序列为 $\boldsymbol{O}_r, r=1,\cdots,r,\cdots,R$,设该语句包含 N_r 个语音段,第 n 个语音段(r,n)的特征序列为 $\boldsymbol{O}_{r,n}$,该语音段给定文本音素为 $q_{r,n}$,MFC 目标函数[13]是 F1 值函数的平滑:

$$F_{\text{MFC}}=\frac{2\sum_{r,n}S(d(r,n))Err(r,n)}{\sum_{r,n}S(d(r,n))+N_W}=\frac{2N_{\text{WW}}^S}{N_D^S+N_W} \tag{3}$$

其中 $N_{\text{WW}}^S=\sum_{r,n}S(d(r,n))Err(r,n)$为平滑的 N_{WW} 值,$N_D^S=\sum_{r,n}S(d(r,n))$为平滑 N_D 值.$Err(r,n)$为语音(r,n)的人工标注结果,错误时 $Err(r,n)=1$,否则为 0.$S(\cdot)$为 Sigmoid 函数:

$$S(u)=\frac{1}{1+\exp(-\theta u)} \tag{4}$$

常系数 $\theta>0$.式(3)中 $d(r,n)$为基于 GOP 的语音段(r,n)的检错测度:

$$d(r,n)=\frac{1}{T_{r,n}}\left[\log\sum_{q\in Q(r,n)}\exp g(q,\boldsymbol{O}_{r,n})-g(q,\boldsymbol{O}_{r,n})\right]+\tau \tag{5}$$

$T_{r,n}$是(r,n)段的时长,$g(q,\boldsymbol{O})$为音素 q 产生观察序列 $\boldsymbol{O}$ 的对数概率的 κ 倍:

$$g(q,\boldsymbol{O})=\kappa\log p(\boldsymbol{O}\mid q) \tag{6}$$

$\kappa\in(0,1)$是减少模型概率范围的因子.式(5)中的 τ 为检错门限,$Q(r,n)$是语音段(r,n)可能的发音假设.由 GOP 原理,$d(r,n)>0$ 时该音素被检测为错误,否则为正确,更多 MFC 定义可见文献[13].

3　基于 MFC 的区分性线性特征补偿

3.1　区域相关线性特征补偿

设 M 为 HMM 集中所有高斯分量的个数,$\boldsymbol{o}_t$ 是 t 时刻的谱特征观察矢量,φ_t^m 是 t 时刻 $\boldsymbol{o}_t$ 在第 m 个分量的后验概率,利用高斯相关的偏移量对原有特征进行补偿形成新的特征 $\boldsymbol{y}_t$:

$$\boldsymbol{y}_t=\boldsymbol{o}_t+\sum_{m=1}^{M}\varphi_t^m\boldsymbol{b}_m \tag{7}$$

$\boldsymbol{b}_m$ 是与 HMM 集中第 m 个高斯分量相关的偏移量.定义偏移量矩阵 $\boldsymbol{B}=[\boldsymbol{b}_1\ \ \boldsymbol{b}_2\ \ \cdots\ \ \boldsymbol{b}_M]$,以及 t 时刻的后验概率矢量 $\boldsymbol{\Phi}_t=[\varphi_t^1\ \ \varphi_t^2\ \ \cdots\ \ \varphi_t^M]^{\text{T}}$,式(7)又可写为:

$$\boldsymbol{y}_t=\boldsymbol{o}_t+\boldsymbol{B}\boldsymbol{\Phi}_t \tag{8}$$

初始条件下 $\boldsymbol{B}=\boldsymbol{0}$.通过 MFC 优化变换 $\boldsymbol{B}$,称为基于 MFC 的区域相关线性补偿(MFC based Region Dependent Linear Compensation,MFC-RDLC).

3.2　参数更新算法

采用无约束优化例程 L-BFGS[17]对 $\boldsymbol{B}$ 进行迭代优化,需计算目标函数值 F_{MFC}及其对 $\boldsymbol{B}$ 的偏导数.目标函数值 F_{MFC}可由式(3)得到,对于偏导数:

$$\frac{\partial F_{\text{MFC}}}{\partial\boldsymbol{B}}=\sum_{t=1}^{T}\frac{\partial F_{\text{MFC}}}{\partial\boldsymbol{y}_t}\frac{\partial\boldsymbol{y}_t}{\partial\boldsymbol{B}}=\sum_{t=1}^{T}\frac{\partial F_{\text{MFC}}}{\partial\boldsymbol{y}_t}\boldsymbol{\Phi}_t^{\text{T}} \tag{9}$$

其中 T 为所有观察帧数.设 y_{ti}是 $\boldsymbol{y}_t$ 的第 i 维,式中目标函数对特征导数计算式为:

$$\partial F_{\text{MFC}}/\partial y_{\text{ti}}=P_{\text{ti}}^{\text{direct}}+P_{\text{ti}}^{\text{indirect}} \tag{10}$$

其中 $P_{\text{ti}}^{\text{direct}}$称直接导数,为均值和方差不变目标函数对特征的导数:

$$P_{\text{ti}}^{\text{direct}}=\frac{\kappa}{T_{r,n}}\sum_q\sum_s\sum_m\psi_{\text{qsmt}}^{\text{MFC}}\frac{\mu_{\text{qsmi}}-y_{\text{ti}}}{\sigma_{\text{qsmi}}^2} \tag{11}$$

其中 $\psi_{\text{qsmt}}^{\text{MFC}}=\psi_q^{\text{MFC}}(r,n)\psi_{\text{qsm}}(t)$,$\psi_{\text{qsm}}(t)$是$(r,n)$中 $\boldsymbol{y}_t$ 在模型 q 状态 s 高斯 m 的后验概率,可前-后向计算得到.$\psi_q^{\text{MFC}}(r,n)$计算式为:

$$\psi_q^{\mathrm{MFC}}(r,n)=$$
$$D(q,r,n)\left(\frac{2}{N_D^S+N_W}E(r,n)-\frac{2N_{\mathrm{WW}}^S}{(N_D^S+N_W)^2}\right) \tag{12}$$

其中

$$D(q,r,n)=\frac{\kappa\theta \mathrm{e}^{-\theta d(r,n)}}{(1+\mathrm{e}^{-\theta d(r,n)})^2}(\psi_q(r,n)-I(q,q_{r,n})) \tag{13}$$

$\psi_q(r,n)$是(r,n)中音素 q 的后验概率. $I(\cdot)$为指示函数,若 q 与 $q_{r,n}$ 相同 $I(q,q_{r,n})$为 1,否则为 0. 关于 $\psi_q^{\mathrm{MFC}}(r,n)$可参见文献[13]. 由于目标函数是模型与特征的函数,又因为模型也是特征的函数,所以导数计算还需考虑特征变化引起模型变化对目标函数的影响,称为间接导数,用 $P_{\mathrm{ti}}^{\mathrm{indirect}}$表示. 对角方差的情况下,设 μ_{qsmi}和 σ_{qsmi}^2为模型 q 状态 s 中高斯 m 的均值和方差的第 i 维,间接导数计算式为:

$$P_{\mathrm{ti}}^{\mathrm{indirect}}=\frac{\partial F_{\mathrm{MFC}}}{\partial \mu_{\mathrm{qsmi}}}\frac{\partial \mu_{\mathrm{qsmi}}}{\partial y_{\mathrm{ti}}}+\frac{\partial F_{\mathrm{MFC}}}{\partial \sigma_{\mathrm{qsmi}}^2}\frac{\partial \sigma_{\mathrm{qsmi}}^2}{\partial y_{\mathrm{ti}}} \tag{14}$$

式中目标函数对模型的偏导数为:

$$\frac{\partial F_{\mathrm{MFC}}}{\partial \mu_{\mathrm{qsmi}}}=\kappa A'_{\mathrm{qsmi}}\sigma_{\mathrm{qsmi}}^{-2} \tag{15}$$

$$\frac{\partial F_{\mathrm{MFC}}}{\partial \sigma_{\mathrm{qsmi}}^2}=\kappa(A''_{\mathrm{qsmi}}\sigma_{\mathrm{qsmi}}^{-4}-A_{\mathrm{qsmi}}\sigma_{\mathrm{qsmi}}^{-2}) \tag{16}$$

其中模型 q 状态 s 高斯 m 的第 i 维零阶、一阶、二阶累积量计算为:

$$A_{\mathrm{qsmi}}=\sum_{t=1}^{T}\psi_{\mathrm{qsmt}}^{\mathrm{MFC}} \tag{17}$$

$$A'_{\mathrm{qsmi}}=\sum_{t=1}^{T}\psi_{\mathrm{qsmt}}^{\mathrm{MFC}}(y_{\mathrm{ti}}-\mu_{\mathrm{qsmi}}) \tag{18}$$

$$A''_{\mathrm{qsmi}}=\sum\psi_{\mathrm{qsmt}}^{\mathrm{MFC}}(y_{\mathrm{ti}}-\mu_{\mathrm{qsmi}})^2 \tag{19}$$

接下来给出式(14)中模型对特征偏导计算式. 在特征优化中,模型采用 ML 更新以避免特征优化通过间接改变模型参数来优化目标函数. 根据 ML 准则,模型对特征的偏导计算式为:

$$\frac{\partial \mu_{\mathrm{qsmi}}}{\partial y_{\mathrm{ti}}}=\frac{\psi_{\mathrm{qsm}}^{\mathrm{ML}}(t)}{\psi_{\mathrm{qsm}}^{\mathrm{ML}}} \tag{20}$$

$$\frac{\partial \sigma_{\mathrm{smi}}}{\partial y_{\mathrm{ti}}}=\frac{\psi_{\mathrm{qsm}}^{\mathrm{ML}}(t)}{\psi_{\mathrm{qsm}}^{\mathrm{ML}}}(y_{\mathrm{ti}}-\mu_{\mathrm{qsmi}}) \tag{21}$$

$\psi_{\mathrm{qsm}}^{\mathrm{ML}}(t)$是 ML 更新前后向计算中 $\boldsymbol{y}_t$ 在模型 q 状态 s 中高斯 m 的后验概率,$\psi_{\mathrm{qsm}}^{\mathrm{ML}}$是 $\psi_{\mathrm{qsm}}^{\mathrm{ML}}(t)$的累加:

$$\psi_{\mathrm{qsm}}^{\mathrm{ML}}=\sum_{t=1}^{T}\psi_{\mathrm{qsm}}^{\mathrm{ML}}(t) \tag{22}$$

计算完目标函数对变换的导数后,调用 L-BFGS 更新变换并重新计算特征. 不断迭代直至收敛. 最后,在测试集上计算新特征并测试,训练过程总结如下:

输入:ML 基线模型,非母语数据及其人工检错标注结果
输出:MFC 最优变换 $\boldsymbol{B}$
开始
Step0 初始化 $\boldsymbol{B}\rightarrow 0$;
Step1 利用基线模型对每帧计算 $\boldsymbol{\Phi}_t$;
Step2 利用基线模型在 L2 数据上进行 ML 更新;
Step3 利用第 2 步得到的模型,对 L2 数据的正确发音段做前向-后向计算获得累积量 $\psi_{\mathrm{qsm}}^{\mathrm{ML}}(t)$及 $\psi_{\mathrm{qsm}}^{\mathrm{ML}}$;
迭代开始
Step4 在训练集上进行第 1 次累加计算,计算各段 GOP 值;
Step5 搜索最佳阈值;
Step6 计算 N_{WW}^S和 N_W^S 及目标函数 F_{MFC};
Step7 进行第 2 次累加,计算 A_{qsmi}、A'_{qsmi}和 A''_{qsmi};
Step8 计算式(14)中所需的目标函数对模型参数的偏导数;
Step9 进行第 3 次累加计算,使用式(10)计算目标函数对各 $\boldsymbol{y}_t$ 的偏导,并用式(9)累加目标函数对变换的导数;
Step10 将目标函数值与导数送入 L-BFGS 更新变换;
Step11 利用新变换重新计算特征;
Step12 进行第 4 次累加计算,利用第 3 步中的前后向数据 $\psi_{\mathrm{qsm}}^{\mathrm{MLE}}(t)$使用 ML 更新公式更新均值与方差;
Step13 若未达到预定迭代次数,返回 Step4;
结束

4 实验与结果

4.1 数据库和实验配置

本文通过面向新疆大学在校预科语言学习的维吾尔族大学生汉语发音检错来验证算法有效性. 基线模型训练自 863 汉语语音库(L1)的 86271 条语句. 谱特征包括 39 维:经过倒谱均值归一化的 13 维美尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)及一阶、二阶差分. 基线模型采用 ML 训练,因 L2 数据有限,采用单音子保证鲁棒性. 模型集包括 67 个 HMM(28 个声母,37 个韵母以及静音与短暂停). 非静音 HMM 均由三状态组成,每个状态为 8 高斯.

MFC 训练在非母语(L2)语音库上进行. 该语音库为 100 名维吾尔族大学生的朗读数据,每人朗读 2 ~ 3 套文本,每套文本包括 50 个单字、25 个词以及 20 个短句并人工标注发音错误. 分为 L2 训练集(18643 句)、L2 测试集(7030 句).

获得基线后,利用该模型对 L2 数据强制对齐获得各音素起止时间,并对各音素(r,n)加入可能的竞争假设形成混淆网络. 若该音素为声母,加入所有的声母模型作为竞争假设,否则加入所有的韵母. 按语音识别的经验设置,式(6)中的 κ 选为 $\kappa=0.1$,式(4)中 θ 选取为 $\theta=10.0$.

4.2 MFC-RDLC 特征训练结果

表 1 给出特征训练的目标函数 F_{MFC}与 $F1$ 值. 先给出利用 MFCC 特征与 ML 训练基线. 在 L2 训练集上搜索

音素相关检错门限(式(5)),直至 F_{MFC}最大,过程可见文献[13],并用该阈值对测试集检错.利用 MFCC 特征与 ML 基线模型和音素独立检错门限,训练和测试集的 $F1$ 值分别为 0.406 和 0.380.

接下来进行 MFC-RDLC 训练,由于基线模型在 L1 数据上使用 ML 训练得到,而 RDLC 训练中假设模型在 L2 数据上采用 ML 训练获得,该假设并不满足.因此我们在 L1 基线系统上增加了一步 L2 数据上的 ML 更新,将更新后的模型作为 RDLC 训练的模型参数的起始点(步骤 2).该模型(表 1 基线 + L2 ML)在训练和测试集 $F1$ 值为 0.388 和 0.381.表明在 L2 集上的 ML 训练对提高检错性能没有帮助,该步骤是为了 RDLC 训练中目标函数平稳单调增长.

从实验看,特征训练迭代 10 次能够达到较好的效果.对于式(8)中的后验概率向量 $\boldsymbol{\Phi}_t$,由于 HMM 集中共 1592 个高斯,其维数为 1592×1.只使用当前时刻后验概率的情况下(CXT = 1),训练和测试集的 $F1$ 值分别为 0.512 和 0.441.较采用原始特征的基线在训练和测试集上的结果(0.406 和 0.380)有明显改进,说明 MFC 特征训练对改进检错性能是有效的.

语音识别中,采用上下文声学特征已证明对识别率提升有帮助,接下来增大上下文窗口进行实验,表 1 中第 2 行(CXT = 3)表示当前帧,前后各一帧的 3 帧后验概率矢量串接起来作为更高维的后验概率矢量,此时后验概率向量维数是 4776×1,训练和测试集上的 $F1$ 值为 0.556 和 0.443. 增加窗口宽度到 CXT = 9,训练集 $F1$ 值上升至 0.598,而测试集 $F1$ 值仅上升至 0.455,训练集 $F1$ 增量远高于测试集,说明引入上下文信息会改进性能,但可训练参数增加也带来了过训练.

表 1　MFC-RDLC 特征参数训练结果

	训练集		测试集	
	F_{MFC}	$F1$	F_{MFC}	$F1$
基线	0.352	0.406	0.344	0.380
基线 + L2 ML	0.367	0.388	0.362	0.381
RDLC CXT = 1	0.497	0.512	0.430	0.441
RDLC CXT = 3	0.537	0.556	0.437	0.443
RDLC CXT = 5	0.551	0.567	0.440	0.447
RDLC CXT = 7	0.567	0.577	0.442	0.452
RDLC CXT = 9	0.584	0.598	0.453	0.455
RDLC AVG	0.585	0.594	0.461	0.465

一种改进办法是采用文献[15]提出的将前 6 – 9 帧、后 6 – 9 帧、前 3 – 5 帧、后 3 – 5 帧、前 1 – 2 帧、后 1 – 2 帧的后验概率矢量平均并与当前 $\boldsymbol{\Phi}_t$ 串接形成高维矢量 . 这种扩展一定程度上避免了短时内奇异点跳变造成后验概率计算不准确[18]. 此时测试 $F1$ 值为 0.465 (RDLC AVG),与直接串接前后 9 帧后验概率矢量(CXT = 9)相比,对过训练有一定的鲁棒性.总体看使用上下文特征能够改进性能,与语音识别中采用上下文特征改进识别性能结论是一致的.从表 1 看,不同窗口大小训练集 F_{MFC}与 $F1$ 值均明显提高,说明 MFC 优化能够增加训练集 $F1$ 值,同时增加训练集 $F1$ 值又能提高测试集 $F1$ 值,与文献[13]中 MFC 模型训练得到的结论也是一致的.

表 2 列出基线与特征优化的精确度和召回率,训练和测试集上的精确度和召回率都有所增加,表明 MFC 优化也同时提高精确度和召回率,根据 $F1$ 值优化特征参数是改进整体性能的有效手段.

表 2　MFC-RDLC 特征训练的精确度与召回率

	训练集		测试集	
模型	精确度	召回率	精确度	召回率
基线	0.403	0.409	0.352	0.415
RDLC AVG	0.672	0.533	0.488	0.445

4.3　模型训练及特征-模型联合训练结果

在优化特征上进行 MFC 模型训练,称为特征-模型的联合训练,通过联合优化获得更好的性能.为进行对比,表 3 先给出原始特征上进行 MFC 模型训练的结果,模型训练以基线模型为起始点,算法见文献[13].训练和测试集的 $F1$ 值分别为 0.611 和 0.469,该结果较基线的改进源于模型优化.表 3 第二行给出联合训练结果,训练和测试集上 $F1$ 值为 0.647 和 0.492,与单独特征训练或单独模型训练相比,$F1$ 值进一步提高.说明采用区分性模型训练,RDLC 特征也优于传统特征,也说明特征训练 $F1$ 值提升并不是通过间接调整模型参数获得的.

表 3　特征参数-模型参数联合训练结果

特征/模型	训练集		测试集	
	F_{MFC}	$F1$	F_{MFC}	$F1$
原始特征/MFC 模型训练	0.562	0.611	0.435	0.469
RDLC 特征/MFC 模型训练	0.628	0.647	0.475	0.492

4.4　区分性特征/模型训练中的其他指标变化

ML 准则和贝叶斯风险最小化(Minimum Bayesian Risk,MBR)准则是语音识别中常见训练准则.接下来考察不同特征与模型条件下两种准则的变化.ML 准则为给定模型对数据的似然度:

$$F_{\mathrm{MLE}}=\sum_{r=1}^{R}\sum_{n=1}^{N_r}\log\left(p^{\kappa}(\boldsymbol{O}_{r,n}\mid q_{r,n})P^{\kappa}(q_{r,n})\right)\quad(23)$$

MBR 目标函数为:

$$F_{\mathrm{MBR}}=\frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N_r}\frac{p^{\kappa}(\boldsymbol{O}_{r,n}\mid q_{r,n})P^{\kappa}(q_{r,n})}{\sum_{q\in Q(r,n)}p^{\kappa}(\boldsymbol{O}_{r,n}\mid q)P^{\kappa}(q)} \tag{24}$$

N 是总音素个数,该函数最大化表示提高音素分类正确度.表 4 给出基线、特征训练、MFC 模型训练以及联合训练中帧平均对数似然度和 F_{MBR}指标.可看到特征训练在训练和测试集上的 $F1$ 值都有增加,尽管模型参数采用 ML 更新,但训练集似然度(-95.7)与基线(-77.3)相比是下降的.再观察 MFC 模型训练、联合训练,都发现 $F1$ 值提高但似然度和 F_{MBR}下降.这表明适用于语音识别的 ML 与 MBR 准则与发音检错的 $F1$ 值最大化并不具有直接相关性.

表 4 参数训练过程中其他指标的变化

	F_{MLE}/T		F_{MBR}	
	训练集	测试集	训练集	测试集
基线	-77.3	-72.3	0.658	0.653
RDLC 特征训练	-95.7	-92.5	0.599	0.600
原始特征 + MFC 模型训练	-82.4	-78.1	0.436	0.433
RDLC 特征 + MFC 模型训练	-96.5	-93.2	0.397	0.396

5 结论

本文将最大化 $F1$ 值准则用于发音错误检测的特征优化,利用区域相关偏移量对谱特征进行线性补偿.推导了最大化 $F1$ 值准则函数对变换的偏导计算式,并利用 L-BFGS 方法进行更新.实验表明该方法能有效提高 $F1$ 值、精确度和召回率.在特征补偿基础上进行区分性模型参数训练,进一步提高了检错性能.该结果验证了基于最大化 $F1$ 准则的区分性特征优化在发音错误检错任务上的有效性.

参考文献

[1] Wei S, Hu G P, Hu Y, Wang R H. A new method for mispronunciation detection using support vector machine based on pronunciation space models[J]. Speech Communication, 2009, 51(10): 896 – 905.

[2] Witt S M, Young S J. Phone-level pronunciation scoring and assessment for interactive language learning[J]. Speech Communication, 2000, 30(2 – 3): 95 – 108.

[3] 葛凤培,潘复平,董滨,颜永红.汉语发音质量评估的实验研究[J].声学学报,2010,35(2):261 – 266.
GE Feng-pei, PAN Fu-ping, DONG Bin, YAN Yong-hong. Experiment investigation of Putonghua quality assessment[J]. Acta Acustica, 2010, 35(2): 261 – 266. (in Chinese)

[4] Lo W K, Zhang S, Meng H. Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system[A]. Proceedings of Interspeech[C]. JAPAN: ISCA, 2010. 765 – 768.

[5] Qian X, Soong F, Meng H. Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)[A]. Proceedings of Interspeech[C]. JAPAN: IEEE, 2010. 757 – 760.

[6] Luo D, Yang X, Wang L. Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus[A]. Proceedings of Interspeech[C]. Italy: ISCA, 2011. 1593 – 1596.

[7] Qian X, Meng H, Soong F. The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training[A]. Proceedings of Interspeech[C]. USA: ISCA, 2012. 775 – 778.

[8] Lee A, Zhang Y, Glass J. Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams[A]. Proceedings of ICASSP[C]. Canada: IEEE, 2013. 8227 – 8231.

[9] Hu W, Qian Y, Soong F. A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)[A]. Proceedings of Interspeech[C]. France: ISCA, 2013. 1886 – 1890.

[10] Juang B H, Katagiri S. Discriminative learning for minimum error classification[J]. IEEE Transactions on Signal Processing, 1992, 40(12): 3043 – 3054.

[11] Bahl L R, Brown P F, Souza P, Mercer R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition[A]. Proceedings of ICASSP[C]. JAPAN: IEEE, 1986. 49 – 52.

[12] Povey D, Woodland P. Minimum phone error and I-smoothing for improved discriminative training[A]. Proceedings of ICASSP[C]. USA: IEEE, 2002. 105 – 108.

[13] Huang H, Wang J, Abudureyimu H. Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning[A]. Proceedings of Interspeech[C]. USA: ISCA. 2012. 815 – 818.

[14] Droppo J, Acero A. Maximum mutual information SPLICE transform for seen and unseen conditions[A]. Proceedings of Interspeech[C]. Portugal: ISCA, 2005. 989 – 992.

[15] Povey D, Kingsbury B, Mangu L, et al. fMPE: discriminatively trained features for speech recognition[A]. Proceedings of ICASSP[C]. USA: IEEE, 2005. 961 – 964.

[16] Zhang B, Matsoukas S, Schwartz R. Discriminatively trained region dependent feature transforms for speech recognition[A]. Proceedings of ICASSP[C]. FRANCE: IEEE, 2006. I313 – I316.

[17] Nocedal J, Wright S J. Numerical Optimization[M]. Germany: Springer, 1999.

[18] 竺博.区分性训练和区分性自适应在自动语音识别声学

模型优化中的应用[D].安徽合肥:中国科学技术大学,2009.

ZHU Bo. Application of discriminative training and discriminative training-adaptation in acoustic modeling of ASR[D]. Hefei, Anhui: University of Science and Technology of China, 2009. (in Chinese)

作者简介

黄　浩　男,1976 年 10 月出生,新疆乌鲁木齐人,副教授.1999 年、2004 年和 2008 年分别在上海交通大学、新疆大学、上海交通大学获工学学士、硕士和博士学位.主要从事语音信号处理、自然语言处理与多媒体交互方面的研究工作.

E-mail:hwanghao@gmail.com

徐海华　男,1975 年 2 月出生,湖北黄冈人.1998 年、2005 年和 2010 年分别在哈尔滨理工大学、华中科技大学、上海交通大学获工学学士、硕士和博士学位.目前为新加坡南洋理工大学 Temasek 实验室研究科学家.主要从事语音识别、关键词检索等方面的研究工作.

E-mail:haihuaxu@ntu.edu.sg

王羡慧　男,1980 年 5 月出生,新疆泽普人,副教授.2004 年、2011 年分别在新疆大学和西安交通大学获工学学士和博士学位.主要从事人工智能与机器学习等方面的研究工作.

E-mail:wisdom@xju.edu.cn

吾守尔·斯拉木　男,中国工程院院士,新疆大学信息科学与工程学院教授.主要研究方向为语音识别、语音合成、多语种信息处理.

E-mail:wushour@xju.edu.cn