声学模型区分性训练中的动态加权数据选取方法

陈 斌 [1] 牛 铜 [1] 张连海 [1] 李弼程 [1] 屈 丹 [1]

摘 要 提出了一种基于动态加权的数据选取方法, 并应用到连续语音识别的声学模型区分性训练中. 该方法联合后验概率和音素准确率选取数据, 首先, 采用后验概率的 Beam 算法裁剪词图, 在此基础上依据候选词所在候选路径的错误率, 基于后验概率动态的赋予候选词不同的权值; 其次, 通过统计音素对之间的混淆程度, 给易混淆音素对动态地加以不同的惩罚权重, 计算音素准确率; 最后, 在估计得到弧段期望准确率分布的基础上, 采用高斯函数形式对所有竞争弧段的期望音素准确率软加权. 实验结果表明, 与最小音素错误准则相比, 该动态加权方法识别准确率提高了 0.61 %, 可有效减少训练时间.

关键词 区分性训练, 语音识别, 训练数据选取, 动态加权

引用格式 陈斌, 牛铜, 张连海, 李弼程, 屈丹. 声学模型区分性训练中的动态加权数据选取方法. 自动化学报, 2014, **40**(12): 2899−2907

DOI 10.3724/SP.J.1004.2014.02899

A Variable Weighting Based Training Data Selection Method for Discriminative Training of Acoustic Models

CHEN Bin[1] NIU Tong[1] ZHANG Lian-Hai[1] LI Bi-Cheng[1] QU Dan[1]

Abstract By combining the phone posterior and phone accuracy, a data selection method based on variable weighting is proposed to improve the discriminative training performance of the acoustic model for continuous speech recognition. Firstly, the word lattice is reduced by using a posterior-based Beam pruning method, and for each hypothesis word a weight is derived from the word error rates of the path containing that word with the posterior. Then, each pair of confusing phones is variably weighted according to a phone confusion matrix, and the modified phone accuracy is calculated by applying those weights. Finally, the distribution of the expected phone accuracies is estimated and all competing arcs are soft weighted using Gaussian functions. Experimental results show that compared with the minimum phone error criterion, the variable weighting method not only improves the recognition rate by 0.61 %, but also reduces the required training time.

Key words Discriminative training, speech recognition, training data selection, variable weighting

Citation Chen Bin, Niu Tong, Zhang Lian-Hai, Li Bi-Cheng, Qu Dan. A variable weighting based training data selection method for discriminative training of acoustic models. *Acta Automatica Sinica*, 2014, **40**(12): 2899−2907

语音识别中常采用最大似然准则进行声学模型的训练, 但最大似然准则只考虑增大正确类别的似然度, 并未考虑错误分类的信息, 可能会造成在提高正确类别相似度的同时, 错误类别的似然度变得更高现象的出现. 而区分性训练不是以最大化训练语料的似然度为目标, 而是关注如何调整不同模型间的分类面, 以提高模型的辨识能力.

近年来, 区分性训练较为明显地提高了语音识别系统的性能, 其中具有代表性的训练准则有最大互信息 (Maximum mutual information, MMI)[1]、最小分类错误 (Minimum classification error, MCE)[2]、最小音素/词错误 (Minimum phone/word error, MPE/MWE)[3]. 为了进一步提高模型的泛化性能和鲁棒性, 类似于机器学习中大决策边距的方法应用到语音识别领域, 以增大正确识别结果与错误识别结果间的距离. 如最大边距估计准则[4]、软边距估计准则[5]、强化混淆信息方法主要有增进的最大互信息/最小音素错误 (Boosted MMI/MPE, BMMI/BMPE)[6] 以及不同模型之间的区分性组合[7−8]. 不少学者将区分性训练方法应用于深层神经网络[9−11] 中, 并得到了性能有效的提升. 由于区分性训练方法同时考虑了正确识别结果和错误候选, 会不可避免地提高算法的运算复杂度.

为提高语音识别系统的实用性, 减少运算量, 需要从大量的训练数据和候选路径中选取有效的数据, 去除没有或是区分性信息很少的数据进行模型的

收稿日期 2013-12-30 录用日期 2014-03-31
Manuscript received December 30, 2013; accepted March 31, 2014
国家自然科学基金 (61175017) 资助
Supported by National Natural Science Foundation of China (61175017)
本文责任编委 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 解放军信息工程大学信息系统工程学院 郑州 450002
1. Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou 450002

训练[12−14]. 根据识别任务, 可以在不同的定义域和不同的单位层次上选取数据, 如特征层、训练语句、词图中的词段落、音素段落或时间帧等. 现有的区分性训练目标准则均在不同的物理意义下, 对训练数据进行了一定的选取. 最大边距估计准则[4] 在似然度定义域中, 通过设立门限值, 定义支持向量集合在训练语句层选取数据. 软边距准则[5] 在似然度定义域中增加了类别信息的比对, 将语句和帧层次的数据单位统一到一个损失函数中, 在这两个单位上选取数据. 而 Liu 等[15] 提出的基于最小音素错误准则的数据选取方法, 是基于状态中高斯分布后验概率的熵值, 在时间帧单位上选取训练样本. 文献 [16] 通过设立门限值范围, 采用一种在期望音素准确率域选取训练语句和候选弧等数据的方法. 但以上数据选取方法与声学模型更新过程所需统计量计算的关联度不够.

为了提高数据选取与统计量的结合度, 提升区分性训练的效率, 本文采用动态加权的方法, 联合后验概率和期望音素准确率域, 进行训练样本和竞争候选的选取, 通过权值的大小来突出训练数据的重要程度, 让区分性训练算法专注于有贡献的样本来调整声学模型参数. 本文首先在后验概率词图中选取对模型统计量贡献较大的候选路径和候选弧, 基于候选路径的错误率和后验概率对候选词加权; 其次, 根据音素对的混淆信息加以惩罚权重, 计算音素准确率, 在此基础上, 估计候选弧段期望音素准确率的分布, 对候选弧加权; 最后, 对本文所提方法的性能进行了讨论.

1 区分性训练方法

1.1 最小音素错误准则

区分性算法一般都采用词图描述候选词之间的竞争关系, 基于词图信息进行声学模型统计量的计算和调整. 词图中包含了许多候选路径, 有效地描述了正确识别结果和候选词之间的竞争关系, 为区分性训练提供了足够的混淆信息[17]. 给定某一句训练语句 z, 最小音素错误准则的目标函数为

$$F_{MPE}(\Lambda)=\sum_{z=1}^{Z}\sum_{s_{zi}\in S}P_{\Lambda}(s_{zi}|X_z)RawAcc(S_{zR},s_{zi}) \tag{1}$$

其中, X_z 为语句 z 的特征向量序列, S_{zR} 为 X_z 对应的正确识别结果, s_{zi} 为语句 z 在词图上的候选序列之一, S_{zR} 根据建模单元和识别任务可以是音素、音节、词和字符串等, 本文以词作为代表进行论述. S 为语音识别器产生的所有可能候选词序列所成集合, $P_{\Lambda}(s_{zi}|X_z)$ 为候选词序列 s_{zi} 的后验概率. $RawAcc(S_{zR},s_{zi})$ 代表正确识别个数减去插入、删除和替换错误个数. 通过对式 (1) 构造辅助函数求解, 可以得到分子、分母项一阶统计量的表达式为

$$\theta_{jm}^{num}(X)=\sum_{z=1}^{Z}\sum_{q\in s_{zi}}\sum_{t=s_q}^{e_q}\gamma_{qjm}(t)\max(0,\gamma_q^{zMPE})X(t) \tag{2}$$

$$\theta_{jm}^{den}(X)=\sum_{z=1}^{Z}\sum_{q\in s_{zi}}\sum_{t=s_q}^{e_q}\gamma_{qjm}(t)\max(0,-\gamma_q^{zMPE})X(t) \tag{3}$$

$$\gamma_q^{zMPE}=\gamma_q(c(q)-c_{avg}^{z}) \tag{4}$$

其中, q 为词图中的一条弧, 其起始时间和结束时间分别为 s_q 和 e_q, c_{avg}^{z} 为词图中所有候选路径的平均音素准确率, $c(q)$ 为词图中候选路径包含弧 q 的期望音素准确率, γ_q 为弧 q 的后验概率, $\gamma_{qjm}(t)$ 为弧 q 在时刻 t 处于状态 j 中第 m 个高斯混元的后验概率.

1.2 强化混淆信息的区分性训练准则

基于强化混淆信息的区分性训练方法进一步提高了识别性能, 其通过对候选路径加权, 使得正确识别结果与错误较多的候选词序列不能离得太近. 即对于每一个训练语句, 在候选词序列空间上, 某候选词序列与正确识别结果差别越大, 则该候选词序列具有较强的区分性信息, 在训练时应该增大其权重, 以便在声学模型调整时得到重视. BMPE 的目标函数为

$$F_{BMPE}(\Lambda)=\sum_{z=1}^{Z}\sum_{s_{zi}\in S}P_{\Lambda}(s_{zi}|X_z)\times RawAcc(S_{zR},s_{zi})\cdot \mathrm{e}^{\sigma\varepsilon(S_{zR},s_{zi})} \tag{5}$$

式中, $\varepsilon(S_{zR},s_{zi})$ 为正确识别结果 S_{zR} 在候选词序列 s_{zi} 上错误识别词的个数, σ 为增进因子, 目的为调整 $\varepsilon(S_{zR},s_{zi})$ 所产生影响, σ 越大越能凸显正确识别结果与错误候选词序列间的差异.

2 基于后验概率的动态加权

基于强化混淆信息的区分性训练方法对每一条候选路径采用相同的加权因子 σ, 并且需要经验设定. 本文采用另一种单元更小的基于后验率的动态加权 (Posterior probability based weighting, PPW), 根据训练语句的识别率和每个候选词的后验概率, 动态给每个候选词加以不同的权值, 避免强化混淆信息目标函数中 σ 的经验设定. 先定义一句

训练语句 z 的识别错误率 E_z:

$$E_z = \frac{1}{2N}\sum_{n=1}^{N_1}\sum_{i=1}^{N_2}\Big(1 - P_\Lambda(S_{zR}|X_z(n)) + P_\Lambda(s_{zi}|X_z(n))\Big) \tag{6}$$

其中, $N = N_1N_2$ 为归一化因子, 使得 $E_z \in [0,1]$, N_1 为该训练语句 z 中正确标注的词个数, N_2 为正确识别结果所对应的竞争候选词个数, S_{zR} 为特征 $X_z(n)$ 所对应的正确识别结果, s_{zi} 为候选的竞争词序列, $P_\Lambda(S_{zR}|X_z(n))$ 是训练语句 z 中特征 $X_z(n)$ 在声学模型 Λ 下正确识别成 S_{zR} 的后验概率, $P_\Lambda(s_{zi}|X_z(n))$ 是训练语句 z 中特征 $X_z(n)$ 在声学模型 Λ 下正确识别成 s_{zi} 的后验概率.

2.1　基于语句识别错误率的动态权重

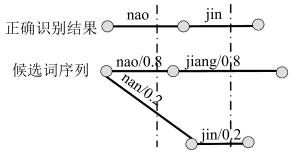
基于识别错误率 E_z 的定义, 对于句子中所有的候选词 s_{zi}, 根据其与正确识别结果 S_{zR} 的后验概率之差来加以不同的权重 $VW_z(S_{zR}, s_{zi})$, 以表示句子 z 的候选词 s_{zi} 对于训练的重要程度.

$$VW_z(S_{zR}, s_{zi}) = \alpha_z^{\frac{1}{2}(1-P_\Lambda(S_{zR}|X_z(n))+P_\Lambda(s_{zi}|X_z(n)))} \tag{7}$$

其中, α_z 是与训练语句 z 识别错误率相关的参数, 其求解式为 $\alpha_z = E_z/(1-E_z)$.

当训练语句的错误率 $E_z < 0.5$ 时, 表示该训练语句中一半以上的词会被正确地识别, 则其 $\alpha_z \in [0,1]$. 又因为 $VW_z(S_{zR}, s_{zi})$ 的指数项 $(1/2)(1 - P_\Lambda(S_{zR}|X_z(n)) + P_\Lambda(s_{zi}|X_z(n))) \in [0,1]$, 因此对于此训练语句中所有词的权重 $VW_z(S_{zR}, s_{zi}) \in [\alpha_z, 1]$. 而当训练语句的错误率 $E_z > 0.5$ 时, 表示该训练语句中一半以上的词都将被错误地识别, 则其 $\alpha_z > 1$. 因此对于此训练语句中所有词的权重 $VW_z(S_{zR}, s_{zi}) \in [1, \alpha_z]$, 权重会大于 1, 此训练语句在训练过程中将会受到重视. 以图 1 为例, 说明本文权重 $VW_z(S_{zR}, s_{zi})$ 的选取, 图中将权重表达式 $VW_z(S_{zR}, s_{zi})$ 简记为 $VW_z(s_{zi})$.

图 1 为一训练语句 z 中的一段语音, 其正确识别结果分别为脑筋/nao/、/jin/和南疆/nan/、/jiang/, 假设词图中只有 2 条候选序列, 其后验概率如图 1 所示. 图 1 (a) 假设该训练语句的错误率 $E_z = 0.3$, 进而求得 $\alpha_z = 0.43$ 和各个候选序列的权值. 根据后验概率值可知, /nao/将正确地识别, 而/jin/将被错误地识别成/jiang/, 但正确候选词/jin/的权重 $VW_z(\text{jin}, \text{jin})$ 会小于错误候选词/jiang/的权重 $VW_z(\text{jin}, \text{jiang})$, 且同样是候选错误词竞争集, $VW_z(\text{jin}, \text{jiang})$ 会大于 $VW_z(\text{nao}, \text{nan})$. 对于 $E_z < 0.5$ 的训练语句, 某个词 S_{zR} 较为可能被识别成错误的候选词 s_{zi} 时, 在声学模型训练中应该重视该词, 即增大它的权值 $VW_z(S_{zR}, s_{zi})$. 这样的权重设计具有强调那些易识别错误的词段落能力. 然而如果一个训练语句的 $E_z > 0.5$, 由于一半以上的词都将被错误识别, 此时应该主要考虑那些正确识别的词所提供的区分性信息, 即在保证较为正确识别词的基础上, 再去提高该词的其他候选词段落的权值, 因此 $VW_z(\text{jiang}, \text{jin})$ 会大于 $VW_z(\text{nan}, \text{nao})$, 如图 1 (b) 所示, 假设该训练语句的错误率 $E_z = 0.7$, 可求得 $\alpha_z = 2.33$.

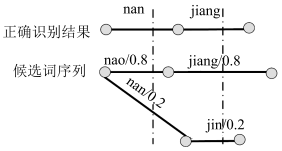


$VW_z(\text{nao}) = 0.43^{0.5} = 0.66,\ VW_z(\text{jiang}) = 0.43^{0.2} = 0.84$

$VW_z(\text{nan}) = 0.43^{0.8} = 0.51,\ VW_z(\text{jin}) = 0.43^{0.5} = 0.66$

(a) 错误率为 0.3 时候选词权重

(a) The weights of hypothesis words when error rate is 0.3



$VW_z(\text{nao}) = 2.33^{0.2} = 1.18,\ VW_z(\text{jiang}) = 2.33^{0.5} = 1.53$

$VW_z(\text{nan}) = 2.33^{0.5} = 1.53,\ VW_z(\text{jin}) = 2.33^{0.8} = 1.96$

(b) 错误率为 0.7 时候选词权重

(b) The weights of hypothesis words when error rate is 0.7

图 1　权重的选择示意图

Fig. 1　The illustration of weight selection

2.2　基于后验概率的词图数据选取

由于词图中含有丰富的信息, 会使在基于后验概率的加权过程中出现的候选路径非常巨大, 其中多数后验概率过小的候选词对统计量的贡献较小, 不仅会明显地增加区分性训练的耗时, 而且会影响识别的性能. 为了提高区分性训练的效率, 需要对词图进行有效裁剪, 选取对训练有用的样本. 常用的裁剪方法主要应用于连续语音识别的解码过程中, 如基于候选路径似然度的 Beam 裁剪方法[18] 和将词图转化为混淆网络的方法[19] 等. 但与语音识别中的解码不同, 区分性训练中常采用一元文法生成词图竞争集, 需要保留候选弧的起止时间, 计算弧的后验概率, 以便对高斯占有率和统计量加权.

而在基于似然度的 Beam 解码过程中, 只扩展累积似然度大于门限值的状态和路径, 低于门限值的状态和路径将会被裁剪, 这会使很多后验概率较大, 具有竞争性和混淆信息的候选词, 由于所在路径的似然得分较低而被去除, 因而基于似然度的裁剪

并不适合于区分性训练, 但 Beam 裁剪算法仍是一种很有效的词图裁剪方法.

由于后验概率对模型参数的更新和识别结果有较大影响, 因此本文采用基于后验概率的 Beam 算法 (Posterior probability based beam, PPBeam) 进行词弧的裁剪, 即将描述声学模型和语言模型得分的似然度词图转化为后验概率词图, 在后验概率词图上进行裁剪. 设 X_1^{T} 为给定的语音特征, q 为词图中起始时间和结束时间分别为 s_q 和 e_q 的一条弧, 根据全概率公式, 弧 q 的后验概率计算方式为:

$$P_{arc}[q_{s_q}^{e_q}] = P(q_{s_q}^{e_q}|X_1^{\mathrm{T}}) = \frac{p(X_1^{\mathrm{T}}|q_{s_q}^{e_q})p(q_{s_q}^{e_q})}{p(X_1^{\mathrm{T}})} = \frac{\sum\limits_h \sum\limits_f p(X_1^{\mathrm{T}}|\phi(h), q_{s_q}^{e_q}, \varphi(f))^{\gamma} p(\phi(h), q_{s_q}^{e_q}, \varphi(f))^{\lambda}}{P(X_1^{\mathrm{T}}) * P(X_{s_q}^{e_q}|q)} \tag{8}$$

式中, $\phi(h)$ 为词图中起始节点到候选弧 q 的所有前置候选路径词序列, $\varphi(f)$ 为侯选弧 q 到结束节点的所有后续候选路径词序列, $p(X_1^{\mathrm{T}}|\phi(h), q_{s_q}^{e_q}, \varphi(f))$、$p(\phi(h), q_{s_q}^{e_q}, \varphi(f))$ 分别为声学模型和语言模型得分, $P(X_1^{\mathrm{T}}) * P(X_{s_q}^{e_q}|q)$ 为词图中开始节点到结束节点所有路径得分, γ、λ 分别为声学模型和语言模型的权重因子, 式 (8) 可以通过前向 – 后向算法得到.

3 基于混淆信息加权的音素准确率

在讨论后验概率的基础上, 由式 (4) 可知, 期望音素准确率 $c(q)$ 也是非常重要的量, 需结合两者计算统计量, 且 $c(q)$ 决定了某一候选弧是作为分母项还是分子项. 另外, 统计量计算时, 需防止奇异样本使得 γ_q^{zMPE} 中某一项过大, 而让该项受到过分的加权和训练. 式 (4) 中平均音素准确率 c_{avg}^z、期望音素准确率 $c(q)$ 均基于音素准确率的计算, 词图中候选词 W_k 的准确率 $RawAcc(W_k)$ 为所含音素的准确率 $PhoneAcc(q)$ 之和, 即 $RawAcc(W_k) = \sum_{q \in W_k} PhoneAcc(q)$. 由于音素准确率是区分性训练准则的基础, 其计算方法对识别结果有较大的影响, 因此在求解期望音素准确率之前, 需要对音素准确率计算方法进行研究, 较为常用的方法有最小音素帧错误准则[20]. 它通过在时间帧层与正确标注相比较, 利用识别结果中的正确帧数来描述准确率; 状态层最小贝叶斯风险准则[21], 其准确率是采用比较状态层的标注得到; 最小散度准则[22] 通过计算识别结果与候选词模型的 Kullback-Leibler 距离来刻画音素准确率. 但这些方法计算得到的音素准确率变化范围过大, 容易使部分样本出现过分重视现象.

在最小音素错误准则中, 音素准确率 $PhoneAcc(q)$ 的计算方法为

$$PhoneAcc(q) = \max_{z_r} \begin{cases} -1 + 2e(q, z_r), & q = z_r \\ -1 + e(q, z_r), & q \neq z_r \end{cases} \tag{9}$$

式中, $e(q, z_r)$ 为音素 q 和正确音素 z_r 的重叠弧长度.

本文基于混淆信息加权 (Confusion information based weighting, CIW) 音素准确率的求解方法是根据音素间不同的混淆程度动态地加以不同的权重, 对易混淆的音素通过权重加以惩罚, 降低其准确率, 以增大在训练过程中对其的重视程度, 提高易混淆音素对的识别率, 进而提升识别性能. 具体方法为先统计各音素对之间的混淆程度, 根据混淆程度大小排序, 基于错误率所占总错误率的比重确定权值. 由表 1 的统计结果可知, 前 100 个易混淆音素对中所含的错误帧数已占所有错误数的 42.27 %, 所含帧数占总数的 61 %, 这些音素对之间的错误, 是总体识别错误的主要部分, 针对这些音素来调整模型, 能较好地提高识别性能. 为提高训练的效率和突出易混淆样本, 不将 K 扩充到全部错误音素对, 同时为了减小权重的变化范围, 以 10 个音素对为单位进行聚类, 加以相同的权值, 其权值的确定方式为每 10 个音素对所含的错误帧数除以前 100 类所含的总错误帧数, 如下式所示:

$$CIWPhoneAcc(q) = \max_{z_r} \begin{cases} -1 + 2e(q, z_r), & q = z_r \\ -1 + \left(1 - \dfrac{n_k}{\sum\limits_k n_k}\rho\right) e(q, z_r), & q \neq z_r, q \in Q_k \\ -1 + e(q, z_r), & q \neq z_r, q \notin Q_k \end{cases} \tag{10}$$

式中, n_K 表示所属 K 类的错误帧数占总错误数的比例, Q_K 为第 K 类的音素集, ρ 为权重因子.

图 2 是音素准确率的计算示意图. 首先, 比对识别候选词与正确识别结果的重叠长度, 在重叠范围内计算音素准确率, 从中选出最大值作为音素的准确率. 加权音素准确率的计算过程与常用音素准确率的计算方法类似, 只是在求取重叠部分准确率时, 需要比对是否为前 K 对音素混淆对, 以确定其权值. 图 2 中/j-in+g/和/j-ing+j/为前 10 个混淆音素对之一, 假设权值为 0.2, 采用式 (10) 可以得到其准确率. 由于静音不计入最后的识别结果统计, 因此求取词弧准确率时, 均未计算静音和短歇音的准确率.

候选语句的准确率为整个候选路径中所含音素的准确率之和; 接着, 基于本文音素准确率计算方法进一步求得期望音素准确率, 与文献 [16] 不同, 这里采用一种软性的数据选取方法, 即在期望音素准确率域上对所有的数据加权 (Expected phone accuracy based weighting, EPAW), 通过权值的大小来突出训练数据的重要程度和减小动态变化范围. 根据数据的实际分布范围和迭代训练过程, 本文采用高斯函数对数据加权. 具体过程为: 首先对 $c(q)-c^z_{avg}$ 归一化, 使得归一化后的 $\varphi(c(q)-c^z_{avg})\in[-1,1]$, 接着在此基础上统计 $\varphi(c(q)-c^z_{avg})$ 的均值和方差, 最后依据得到的高斯函数对 $\varphi(c(q)-c^z_{avg})$ 项加权.

$$\varphi(c(q)-c^z_{avg})=\frac{c(q)-c^z_{avg}}{N_1} \tag{11}$$

$$SoftW(x)=\frac{1}{C}\mathrm{e}^{-\frac{(x-\mu)^2}{2\delta^2}} \tag{12}$$

在采用上述后验概率和混淆信息加权方法的基础上, 根据 MPE 准则中辅助函数的构造过程, 可以得到本文动态加权区分性训练方法参数求解的辅助函数:

$$\begin{aligned}H_{VW}(\lambda,\lambda')=&\sum_z\sum_q\sum_m\sum_{t=s_q}^{e_q}VW_z(q)\times\\&\gamma^z_q(c(q)-c^z_{avg})SoftW(\varphi(c(q)-c^z_{avg}))\times\\&\gamma^z_{qm}(t)\lg N\Big(X_z(t),\mu_{qm},\Sigma_{qm}\Big)\end{aligned} \tag{13}$$

表 1　前 K 个音素对混淆信息及其权重

Table 1　The top K confusion phone pairs and their weights

K 值	错误帧数占总错误数比例 (%)	权值
1～10	11.77	27.84
11～20	6.79	16.06
21～30	4.88	11.54
31～40	3.56	8.42
41～50	3.25	7.69
51～60	2.95	6.98
61～70	2.62	6.20
71～80	2.37	5.61
81～90	2.13	5.04
91～100	1.95	4.61

4　测试评估

4.1　实验语料库及设置

本节将本文动态加权数据选取的方法应用到连续语音识别中. 实验语料采用中文微软语料库 Speech Corpora (1.0 版本), 其全部语料在安静办公室环境下录制, 采样率为 16 kHz, 16 bit 量化. 训练集共有 19 688 句, 共 454 315 个音节, 测试集共 500 句, 训练集与测试集中语音时长分别约为 33 小时和 0.7 小时.

对语料进行统计, 共有 404 个无调音节, 文中选择声韵母作为模型基元, 零声母 (_a、_o、_e、_i、_u、_v), 加上静音 (sil) 以及常规的声韵母, 一共有 69 个模型基元, 在此基础上将模

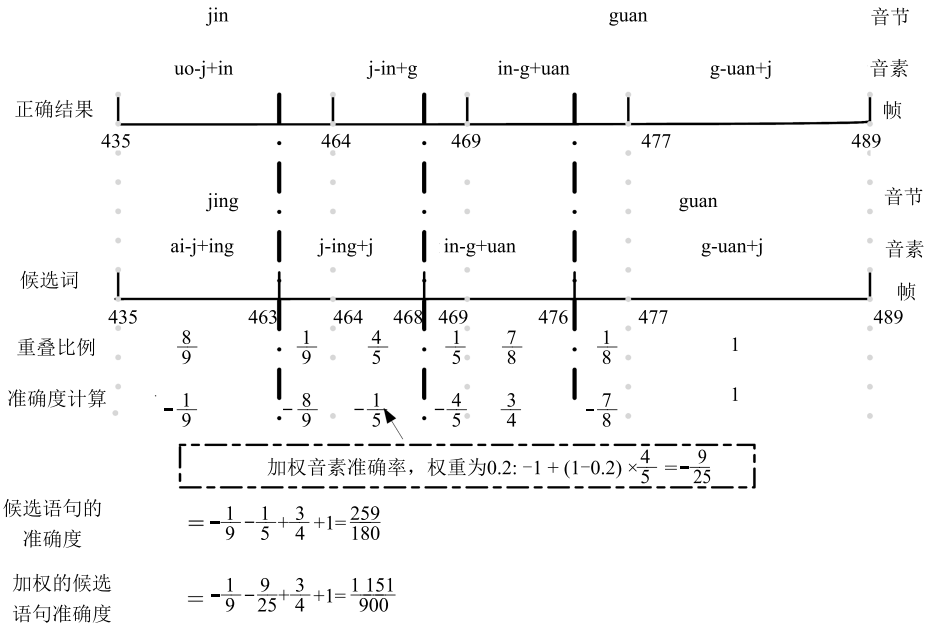


图 2　音素准确度的计算示意图

Fig. 2　The illustration of phone accuracy calculation

型基元扩展为上下文相关的交叉词三音子基元, 简称为三音子.

基于 HTK 3.4 建立基线系统, 声学模型采用 3 状态的自左至右无跨越的隐马尔科夫模型来建模声韵母单元, 在三音子扩展时, 根据前后所有可能声韵母进行扩展, 通过决策树对三音子模型进行状态绑定, 解决数据稀疏的问题, 最终绑定后的模型有效状态数为 2 843 个. 其中单音子模型的训练是通过计算全局均值与方差作为全局初始模型, 采用隐马尔科夫模型参数嵌入式重估算法, 训练所有单音子基元, 然后添加静音模型和短歇音模型, 重新训练模型; 最后对齐训练语料, 获取语料的最优标注, 再次训练模型. 在训练完单音子模型后, 将单音子扩展为上下文相关的三音子模型; 设定问题集, 利用脚本生成决策树, 根据决策树执行状态绑定后重估 4 遍, 从而得到跨词边界的上下文相关三音子模型; 最后, 加入短静音模型, 将其与静音模型的第三个状态共享. 利用 SRILM 工具根据语料库中自有的标注文件训练得到语言模型.

采用准确率进行实验结果的评估, 假设测试语料实际词总数为 N, 正确识别的词总数记为 H, 删除错误的词总数记为 D, 插入错误的词总数记为 I, 替代错误的词总数记为 S, 准确率 A 定义如下:

$$A = \frac{H - I}{N} \times 100\,\% \tag{14}$$

4.2 基于后验概率词图选取实验

本文先得到每种加权方法的识别性能, 最后讨论联合各种方法的识别率. 基于似然度的 Beam 词图裁剪过程为, 先设定一个 Beam 宽度, 各个时刻所有路径累积对数似然度的最大值与 Beam 的宽度之差作为下界, 候选路径中对数似然度大于此下界将保留继续下一时刻的搜索, 否则就会被摒弃. 由于区分性训练过程中, 音素准确率的计算是在正确识别结果的词边界范围内进行识别结果的比对, 因此与运用于整条候选路径上基于似然度的 Beam 词图裁剪方法不同, 本文采用基于后验概率的 Beam 方法, 主要运用于候选弧上, 即在后验概率词图中, 得到某一时刻所有候选弧后验概率的最大值, 将此后验概率值除以 Beam 之商作为门限值, 当前时刻后验概率值小于门限值的候选弧将被裁剪, 如果某个节点的所有弧都被裁剪, 这个节点也将被裁剪, 裁剪示意图如图 3 所示. 区分性训练时采用一元文法, 则 $p(\phi(h), q_{s_q}^{e_q}, \varphi(f)) = p(q_{s_q}^{e_q})$, 经过实验得到 $\gamma = 1/50$, $\lambda = 1/3$.

Beam 宽度值对结果影响较大, 设置太小会使正确的候选词被过早地去除, 设置太大会增大运算量, 减慢解码和后端的处理速度. 文中先采用基于似然度的 Beam (Beam 为 1 000) 词图裁剪方法进行识别译码, 得到基线的词图; 在此基础上, 采用基于后验概率的 Beam 裁剪算法剪枝词图, 讨论其性能. 表 2 为不同的 Beam 宽度下, 基于对数似然度的 Beam (Log likelihood beam, LLBeam) 裁剪方法和基于后验概率 Beam 裁剪方法的词图识别率. 表 3 为不同的 Beam 宽度下, 两种方法所得到的词图规模大小, 其中词图规模用每种词图在硬盘中占用的存储空间来描述.

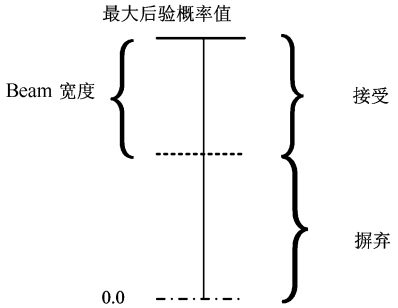


图 3 Beam 裁剪算法

Fig. 3 The Beam algorithm

表 2 不同 Beam 值下的词图识别率 (%)

Table 2 The lattice word accuracy with different Beam (%)

Beam	LLBeam	PPBeam
80	72.45	74.37
90	73.41	74.74
100	74.52	75.89
150	75.18	76.11
200	75.84	76.19
250	76.11	76.19
300	76.18	76.22
400	76.18	76.19
500	76.16	76.22
1 000	76.21	76.22

由表 2 和表 3 可知, Beam 值对识别性能和词图大小有较大的影响, 两种裁剪方法得到的词图会随 Beam 值增大而显著增大. 当 Beam 值较小时, 基于后验概率 Beam 方法的词图会大于基于似然度的 Beam 裁剪方法, 且识别率也会高于相同 Beam 值下基于似然度的 Beam 裁剪方法. 其原因是因为后验概率的动态范围相对较小, 当 Beam 值较小时, 保留的都是后验概率较大、对区分性训练统计量计算贡献较大的候选词弧. 当 Beam 宽度大于 250 时, 两者的识别性能相当, 但基于似然度 Beam 方法的词图会大于基于后验概率的方法. 随着 Beam 值的进一步增大, 两种方法的识别性能并没有明显的变化. 基于后验概率 Beam 方法达到最好性能时 (Beam

为 200) 的词图大小, 为基于似然度方法词图达到最好性能 (Beam 为 250) 词图大小的一半. 这说明适当地去除一些后验概率较低、对统计量贡献较小的候选词和节点, 减小词图尺寸, 对识别结果影响不大. 同时可以降低运算量, 提高区分性训练的效率. 基于后验概率动态加权与期望音素准确率的识别性能在第 4.4 节讨论.

表 3　不同 Beam 值下的词图规模大小

Table 3　The lattice size with different Beam

Beam	LLBeam	PPBeam
80	65.1 M	98.4 M
90	98.5 M	163.8 M
100	148.6 M	1.10 G
150	1.3 G	3.91 G
200	11.63 G	9.97 G
250	22.3 G	16.2 G
300	31.4 G	29.4 G
400	36.1 G	39.6 G
500	39.6 G	42.1 G
1 000	53.2 G	52.1 G

4.3　基于混淆信息加权的音素准确率计算实验

基于混淆信息加权和采用 MPE 中音素准确率计算的识别结果如图 4 所示. 由图 4 可知, 本文准确率的计算方法能有效地提高识别性能, 在迭代次数较高时, 性能的提高更为明显. 说明通过突出和提高易混淆音素对的识别率, 能提高整体的识别性能, 表明了采用混淆信息加权音素准确率计算的有效性.

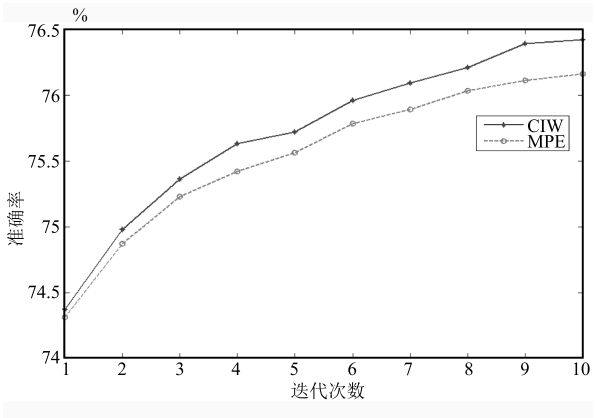


图 4　基于混淆信息加权和 MPE 音素准确率计算的识别性能

Fig. 4　The performance of phone accuracy calculation based on confusion information weighting and MPE

4.4　基于 γ_q^{zMPE} 动态加权的识别实验

词弧后验概率 γ_q 和 $c(q)-c_{avg}^z$ 两部分组成 γ_q^{zMPE}, 共同决定了其对统计量的贡献大小, 对数据统计可得归一化期望音素准确率的变化范围 $-0.84 \le \varphi(c(q)-c_{avg}^z) \le 0.09$, 表明候选弧的期望音素准确率大多都低于平均音素准确率, 基于似然训练的模型能够取得较好的识别效果, 主要需要提高具有竞争性、识别错误候选弧的识别性能, 其加权函数如图 5 所示. 从图 5 可以看出, 本文方法可以依据数据的分布动态地选取加权区域, 侧重于对数据分布较为密集的区域加权. $\varphi(c(q)-c_{avg}^z)=0$ 附近的权值较大, 这主要是因为其位于分子、分母的决策边界处, 对识别结果和统计量的计算有较大的影响.

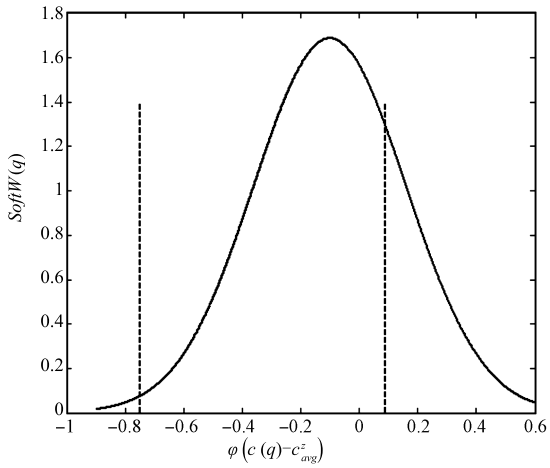


图 5　基于期望音素准确率的加权函数

Fig. 5　The weighting function based on expected phone accuracy

对后验概率、期望音素准确率以及两者联合加权后的识别结果如表 4 所示. 本文的三种加权方法均能够较为明显地加快模型的收敛速度, 同时可以提高识别性能. 其中, 基于后验概率加权方法识别性能提高的幅度会略大于基于期望音素准确率的加权方法. 基于后验概率的加权方法能够随着准确率和后验概率值的大小自动进行权值的选取, 让区分性训练算法越来越致力于易错候选词的训练和识别. 增大正确识别结果与错误较多的候选词序列的距离, 可以有效地避免权重的经验设定, 造成模型的过训练, 提高模型的泛化性能. 同时后验概率和期望音素准确率的联合加权方法可以有效地克服奇异样本对模型的影响, 避免 γ_q^{zMPE} 中某一项过大而使该项受到过分的加权和训练. 另外, 本文算法均是在个人计算机上实现, 一次迭代训练约需要 11 小时; 联合加权后, 可以减少 5 次迭代, 可以减少约两天的训练时间.

4.5　联合基于后验概率动态加权和混淆信息加权实验

实验结果如图 6 所示. 首先基于后验概率的

Beam 算法裁剪词图, 在裁剪后的词图上, 根据其后验概率对候选弧动态加权 (PPBeam+PPW); 然后讨论结合基于音素对混淆度信息的加权 (PPBeam+PPW+CIW); 最后讨论期望音素准确率的加权 (PPBeam+PPW+CIW+EPAW), 并与文献 [16] 方法的识别性能进行了对比.

表 4 基于 γ_q^{zMPE} 加权的识别结果 (准确率 (%))

Table 4 The recognition accuracy upon γ_q^{zMPE} weighting (accuracy rate (%))

迭代次数	MPE	后验概率加权	期望音素正确率加权	两者联合加权
1	74.31	74.46	74.40	74.59
2	74.87	75.13	74.98	75.13
3	75.23	75.81	75.34	75.61
4	75.42	76.09	75.75	76.17
5	75.56	76.28	76.26	76.55
6	75.78	76.42	76.35	76.56
7	75.89	76.41	76.33	76.56
8	76.03	76.38	76.36	76.57
9	76.11	76.40	76.34	76.56
10	76.16	76.40	76.34	76.57

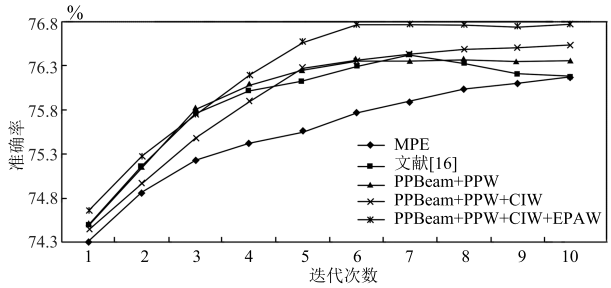


图 6 组合不同数据选取和加权方法的识别性能

Fig. 6 The recognition accuracy by combining different data selection and weighting methods

由图 6 可知, 随着加权方法的组合, 识别性能是提升的, 将所有方法组合一起的 PPBeam+PPW+CIW+EPAW 方法能达到最高的识别性能, 与 MPE 准则和文献 [16] 相比, 识别准确率分别提高了 0.61 % 和 0.55 %, 说明本文数据选取和加权方法具有较好的互补性, 同时, 组合后的方法在第 6 次就可以达到收敛. 而文献 [16] 中的方法在达到最佳识别性能后, 识别率又会随着迭代次数的增加而小幅下降, 这主要是由于过训练引起的, 其所有的数据均在一个域中选取, 数据间选取有一定的重叠性, 且其采用硬性的数据选取方法, 将不在门限范围内的数据直接去除. 因此通过从不同层面和域中选取数据, 根据句子的识别率和训练过程动态地对数据加权, 可以让区分性训练算法更专注于有用和具有竞争性的数据, 在提高模型训练效率的同时不损失识别性能.

5 结论

本文提出了一种联合多域进行数据选取和加权的方法, 并将其应用到连续语音识别的声学模型区分性训练中, 在采用后验概率 Beam 算法对词图裁剪后, 分别基于后验概率、音素对混淆信息和期望音素准确率在区分性训练过程中, 对候选词进行动态加权, 突出对区分性声学模型训练有用的样本. 实验结果表明, 本文方法可以有效加快模型的收敛速度, 提高了模型的泛化性能. 由于 MMI、MPE、MCE 等区分性训练准则可以进行统一表示, 后续将进一步研究本文的数据选取方法与其他的区分性训练准则相结合.

References

1 Valtchev V, Odell J J, Woodland P C, Young S J. MMIE training of large vocabulary recognition systems. *Speech Communication*, 1997, **22**(4): 303−314

2 Juang B H, Chou W, Lee C H. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1997, **5**(3): 257−265

3 Povey D, Woodland P C. Minimum phone error and i-smoothing for improved discriminative training. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Florida, USA: IEEE, 2002, **1**: 105−108

4 Sha F. Large Margin Training of Acoustic Models for Speech Recognition [Ph. D. dissertation], University of Pennsylvania, USA, 2007.

5 Li J Y. Soft Margin Estimation for Automatic Speech Recognition [Ph. D. dissertation], Electrical and Computer Engineering, Georgia Institute of Technology, USA, 2008.

6 Povey D, Kanevsky D, Kingsbury B, Ramabhadran B. Boosted MMI for model and feature-space discriminative training. In: Proceedings of the 2008 International Conference on Acoustics, Speech, and Signal Processing. Las Vegas, USA: IEEE, 2008. 4057−4060

7 Wu Ya-Hui, Liu Gang, Guo Jun. Research on model combination based on model confusion. *Acta Automatica Sinica*, 2009, **35**(5): 551−555
(吴娅辉, 刘刚, 郭军. 基于模型混淆度的模型组合算法研究. 自动化学报, 2009, **35**(5): 551−555)

8 Huang Hao, Li Bing-Hu, Wushour Silamu. Discriminative model combination using decision tree based phonetic context modeling. *Acta Automatica Sinica*, 2012, **38**(9): 1449−1458
(黄浩, 李兵虎, 吾守尔 · 斯拉木. 区分性模型组合中基于决策树的声学上下文建模方法. 自动化学报, 2012, **38**(9): 1449−1458)

9 Seltzer M L, Droppo J. Multi-task learning in deep neural networks for improved phoneme recognition. In: Proceedings of the 2013 International Conference on Acoustics, Speech, and Signal Processing. Vancouver, Canada: IEEE, 2013. 6965−6969

10 Kingsbury B, Sainath T N, Soltau H. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization. In: Proceedings of the 13th Annual Conferece of the International Speech Communication Association. Portland, USA: ISCA, 2012.

11 Veselý K, Ghoshal A, Burget L, Povey D. Sequence-discriminative training of deep neural networks. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013. 2345−2349

12 Toth L. Phone recognition with deep sparse rectifier neural networks. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vancouver, Canada: IEEE, 2013. 6985−6989

13 Vinyals O, Deng L. Are sparse representations rich enough for acoustic modeling? In: Proceedings of the 13th Annual Conference of the International Speech Communication Association. Portland, USA: ISCA, 2012.

14 Zhang W B, Fung P. Discriminatively trained sparse inverse covariance matrices for low resource acoustic modeling. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013. 2350−2354

15 Liu S H, Chu F H, Lin S H, Lee H S, Chen B. Training data selection for improving discriminative training of acoustic models. In: Proceedings of the 2007 IEEE Workshop on Automatic Speech Recognition & Understanding. Kyoto, Japan: IEEE, 2007. 284−289

16 Chen B, Liu S H, Chu F H. Training data selection for improving discriminative training of acoustic models. *Pattern Recognition Letters*, 2009, **30**(13): 1228−1235

17 Qin L, Rudnicky A. The effect of lattice pruning on MMIE training. In: Proceedings of the 2010 International Conference on Acoustics, Speech and Signal Processing. Dallas, USA: IEEE, 2010. 4898−4901

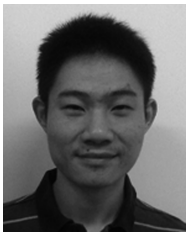
18 Liu Y, Harper M P, Johnson M T, Jamieson L H. The effect of pruning and compression on graphical representations of the output of a speech recognizer. *Computer Speech and Language*, 2003, **17**(4): 329−356

19 Mangu L, Brill E, Stolcke A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 2000, **14**(4): 373−400

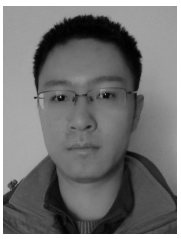
20 Zheng J, Stolcke A. Improved discriminative training using phone attices. In: Proceedings of the 2005 European Confidences Speech Communication and Technology. Lisbon, Portugal: DBLP, 2005. 2125−2128

21 Povey D, Kingsbury B. Evaluation of proposed modifications to MPE for large scale discriminative training. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu. HI: IEEE, 2007. 321−324

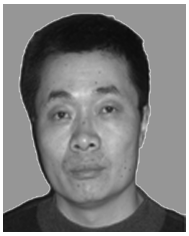
22 Du J, Liu P, Jiang H, Soong F K, Zhou J L, Wang R H. A new minimum divergence approach to discriminative training. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing. Honolulu, HI: IEEE, 2007. IV-677−IV-680



陈 斌 解放军信息工程大学信息系统工程学院博士研究生. 主要研究方向为连续语音识别, 区分性训练. 本文通信作者. E-mail: chenbin873335@163.com
(**CHEN Bin** Ph. D. candidate at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers continuous speech recognition and discriminative training. Corresponding author of this paper.)



牛 铜 解放军信息工程大学信息系统工程学院博士研究生. 主要研究方向为语音增强, 语音识别.
E-mail: niutong0072@gmail.com
(**NIU Tong** Ph. D. candidate at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers speech enhancement and speech recognition.)



张连海 解放军信息工程大学信息系统工程学院副教授. 主要研究方向为语音信号处理, 语音编码与语音识别.
E-mail: lianhaiz@sina.com
(**ZHANG Lian-Hai** Associate professor at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers speech signal processing, speech coding and speech recognition.)

李弼程 解放军信息工程大学信息系统工程学院教授. 主要研究方向为文本分析与理解, 语音处理与识别, 图像/视频处理与识别, 信息融合.
E-mail: lbclm@163.com
(**LI Bi-Cheng** Professor at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers text analysis and understanding, speech/image/video processing and recognition, and information fusing.)

屈 丹 解放军信息工程大学信息系统工程学院副教授. 2005 年获解放军信息工程大学博士学位. 主要研究方向为语音信号处理与模式识别.
E-mail: qudanqudan@sina.com
(**QU Dan** Associate professor at the Institute of Information System Engineering, PLA Information Engineering University. She received her Ph. D. degree from PLA Information Engineering University in 2005. Her research interest covers speech signal processing and pattern recognition.)