

基于联合得分的连续语音识别确认方法*

严斌峰⁺, 朱小燕

(清华大学 智能技术与系统国家重点实验室,北京 100084)

Continuous Speech Recognition and Verification Based on a Combination Score

Yan Bin-Feng⁺, Zhu Xiao-Yan

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: 86-10-62779193, E-mail: ybf99@mails.tsinghua.edu.cn

<http://www.lits.tsinghua.edu.cn/>

Received 2002-12-14; Accepted 2003-04-01

Yan BF, Zhu XY. Continuous speech recognition and verification based on a combination score. *Journal of Software*, 2003,14(12):2014~2020.

<http://www.jos.org.cn/1000-9825/14/2014.htm>

Abstract: In this paper, a speech recognition and verification method is presented based on the integration of likelihood and likelihood ratio. Speech recognition and verification is unified in one-phase framework. In the process of decoding, likelihood ratio is combined with likelihood to get the combination score for searching the final results. Experimental results show that the false-alarm rate is significantly reduced, with only a slight loss in the accuracy rate.

Key words: likelihood ratio test; alternative model; utterance verification; speech recognition

摘要: 提出了一种基于联合概率似然得分和概率似然比得分进行语音识别确认的方法,计算搜索路径得分过程中同时考虑概率似然比得分,在给出系统最终识别结果的同时给出置信度水平.实验结果表明,该方法在大大降低识别系统误警率的同时,基本保持识别正确率不变.

关键词: 似然比检验;备择模型;语音确认;语音识别

中图法分类号: TP391 文献标识码: A

近几年来,随着自动语音识别(automatic speech recognition,简称 ASR)技术的发展,语音识别在对话系统、语音控制系统等领域的应用越来越广泛.在实际应用中,系统必须能够处理不同环境下(例如噪声环境、电话语音等)不同用户自然输入的语音.

实用系统的性能主要取决于对话策略的灵活性和语音识别器的良好性能.然而,传统的连续语音识别系统无法使上述两方面同时达到最优,灵活的对话策略要求识别器必须基于大词表和复杂语法,这使得识别器的搜索空间变大、搜索时间增加、识别准确率降低.特别是在处理自然口语输入的语音时,由于用户语言中可能含

* Supported by the National Natural Science Foundation of China under Grant No.60272019 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G199803050703 (国家重点基础研究发展规划(973))

第一作者简介: 严斌峰(1977-),男,江西鹰潭人,博士生,主要研究领域为语音识别.

有简略和语法错误的语句以及咳嗽声等词表外词汇,要求识别器采用的语法必须是无限制的,识别器的性能会受到更大的影响.有些系统,例如 Rubio^[1]通过系统驱动(system-driven)的严格对话控制保证识别器获得良好的性能,但人机之间的交互(human-machine communication)却不能流畅地进行.

Wilpon^[2],Rose^[3]等人设计了关键词检出(keyword spotting,简称 KWS)系统,可以在对话策略的灵活性和识别器的性能之间取得很好的平衡.传统语音识别系统的目标是准确识别用户语音中的每一个词,而关键词检出系统只识别出用户语音中表达语义的关键词汇,关键词汇表可以根据系统所应用的领域确定.关键词检出系统在检出关键词的同时,必须拒识非关键词,拒绝不含有效关键词和置信度较低的语音段.通过对语音进行确认可以实现对非关键词有效的拒识.

早期 Sukkar^[4]等人提出了模块化的二阶段确认方法.二阶段方法分语音识别和语音确认两个步骤.在语音识别阶段,识别器给出概率似然值得分最高的待确认结果.语音确认阶段对给出的待确认的语音段进行确认,给出置信度水平.其缺点是,语音识别阶段给出待确认结果的语音切分边界并不准确,将严重影响到第 2 个阶段对语音的确认^[4-6].另外,系统在时间上的开销也相应增加.

Rahim^[7]为了克服这个缺点提出了基于统计假设检验的一阶段确认方法,试图将识别步骤和确认步骤合并,计算代表置信度水平的概率似然比得分作为搜索路径得分,识别器给出的结果正是置信度水平最高的识别结果.在训练过程中,以最小分类误差训练方法(minimum classification error,简称 MCE)取代通常的最大似然(maximum likelihood,简称 ML)训练方法,训练目标是概率似然比最大,即确认误差最小(minimum verification error,简称 MVE).一阶段方法的目标是在识别过程中,取得检出性能最优和拒识性能最优两者之间的一个折衷:对关键词的检出性能没有两阶段方法好;相反,对于非关键词的拒识性能却优于两阶段方法.

本文提出一种新的一阶段的识别确认方法:在训练阶段,仍以对训练样本的概率似然值的最大化作为训练目标.在搜索过程中,对每一帧语音特征向量,计算其在相应状态上的概率似然值和概率似然比值,以这两个值的加权和作为该语音特征向量在这个状态上的得分;最后搜索得到的是似然度较高、置信度也较高的识别结果.该方法结合了二阶段模块化方法关键词正确检出率高和传统一阶段方法非关键词正确拒识率高的优点,提高了系统的效率,避免了二阶段方法在初步识别分段时的偏差对最后正确检出的严重影响;同时,对关键词的正确检出率大大高于传统一阶段方法,对非关键词的正确拒识率仅有很小的影响.

在一阶段语音确认方法中,概率似然比的计算非常重要,其关键之处在于备择模型(alternative model)的选择.文献[9]使用音素模型的 N-best 平均分数作为备择模型得分提高了识别性能;文献[8]以动态评价方法与反关键词(anti-keyword)模型结合得到的值作为最后的备择模型得分;文献[7]以与某关键词最相似的几个关键词的训练样本训练该关键词的备择模型.

本文借鉴模型参数共享的思想,以基于分散度的距离标准,采用改进的合并分级聚类算法对词表中所有关键词进行聚类,以某一关键词所在类中其他所有关键词对应的语音为训练样本,训练出该关键词对应的备择模型.基于分散度距离标准的选择可以更充分地利用有限的训练样本,对合并分级聚类算法的改进可以避免分类结果中样本数过于不均的问题,训练出的备择模型相对于文献[7~9]更加精确.

本文第 1 节介绍备择模型的训练.第 2 节介绍识别确认方法.第 3 节是实验结果和分析.第 4 节是结论.

1 备择模型

1.1 统计假设检验

在统计学中,统计假设检验是应用比较广泛的判别准则之一,统计假设检验的基本原理就是比较零假设 H_0 和备择假设 H_1 所得值的大小差异.在语音确认中,

H_0 :表示词 KW_k 存在于一段语音 O 中,且被正确识别;

H_1 :表示词 KW_k 不存在于该段语音中,或存在但被错误地识别为 $KW_l (l \neq k)$.

如果零假设和备择假设都已明确地得到,根据 Neyman-Pearson 引理,最佳的检验通常是概率比检验.如下面的公式所示:

$$PR(k) = \frac{P_k(O|H_0)}{P_k(O|H_1)} \begin{cases} > \tau_k, \text{接受零假设 } H_0 \\ \leq \tau_k, \text{接受备择假设 } H_1 \end{cases} \quad (1)$$

其中 τ_k 是检验的临界阈值。

通常采用似然得分进行评价,此方法称作似然比检验(likelihood ratio test).在实际应用中使用对数似然值,决策规则表示为

$$LR(k) = \log P_k(O|H_0) - \log P_k(O|H_1) = \log P_k(O|\lambda_k) - \log P_k(O|\lambda_k^c) \begin{cases} > \tau_k, \text{接受零假设 } H_0 \\ \leq \tau_k, \text{接受备择假设 } H_1 \end{cases} \quad (2)$$

其中,如果识别过程采用的是 HMM 模型,则 λ_k 为词 KW_k 的 HMM 模型, λ_k^c 为 KW_k 的备择模型.似然比检验方法的关键之处在于备择模型的选择,备择模型的好坏将会严重影响到系统的确认性能,尤其是在区分声学特征相似的关键词时,通过精确的备择模型可以计算得到更大的似然比分值,减少识别结果混淆度。

1.2 备择模型的训练

借鉴模型参数共享的思想,本文提出了基于数据驱动的改进合并分级聚类算法的备择模型训练方法.首先,通过聚类对所有的关键词模型(HMM)进行分类.然后,对某一关键词,找出与其同类的其他所有关键词对应的语音样本,训练出该关键词的备择模型.聚类方法中最重要的两个方面是关键词模型间距离的度量和聚类算法的选择。

描述关键词模型间距离的度量标准,通常利用模型参数,即高斯码本的均值和方差,进行计算得到样本间的距离度量.其优点是实现非常简单,计算复杂度小,且能有效地刻画两个高斯分布之间的相似程度.本文聚类所采用的距离度量为基于分散度的距离度量标准,充分利用有限的训练数据,更能反映数据的声学特性.该距离可以表示为

$$d(i, j) = \frac{1}{N_i + N_j} \left(\sum_k \log \left[\frac{P(O_{ik} | \lambda_i)}{P(O_{ik} | \lambda_j)} \right] + \sum_k \log \left[\frac{P(O_{jk} | \lambda_j)}{P(O_{jk} | \lambda_i)} \right] \right) \quad (3)$$

在上式中, λ_i, λ_j 代表两个不同的声学模型, N_i, N_j 代表各自的训练样本个数, O_{ik} 是训练模型 λ_i 的第 k 个样本, $P(O_{ik} | \lambda_i)$ 表示模型 λ_i 产生样本 O_{ik} 的概率.该度量满足空间距离的非负性、自反性和对称性必要条件,可以作为模型间距离的度量标准。

我们采用快速、实用的合并分级聚类(agglomerative hierarchical clustering)算法,在聚类初始时以每个待分类样本作为一个类别,每次循环时找到相似程度最高的两个类,合并成新类,类别数目不断减少,从而产生一序列新的分类.这种方法的缺点是,在聚类过程中只合不分,不易于控制类内的样本个数,可能导致各个类别的样本数目分布不均,我们对合并分级聚类算法进行适当的改进.在聚类过程中增加必要的调节手段,引入参数 N_{\max}, N_{\min} 用于控制类内样本数的上、下限,对于一个类别的集合 $\{I_i | i \in I\}$, 其中 I_i 代表某一类别, I 为类别标号的集合,作如下操作:

第 1 阶段,在集合中找到一对满足条件

$$d(i, j) = \min_{\forall p, q \in I} \{d(p, q) | N_p + N_q \leq N_{\max}\}, i < j$$

的类别 $I_i, I_j, d(i, j)$ 为这两类间的相互距离, N_p, N_q 为类别 I_p, I_q 各自的样本个数,将 I_j 并入 I_i , 判断是否达到某一收敛标准,若满足则聚类过程停止,否则重复上述步骤.当该阶段的聚类结束之后,得到样本集合的初始划分,并且各个类别的样本数目均限制在 N_{\max} 之内。

第 2 阶段,对初始聚类的结果进行后处理,即二次聚类.对于其中样本数小于 N_{\min} 的类别的集合 $\Psi = \{I_k | N_k < N_{\min}, k \in K\}, K \subset I$, 找到一对满足条件

$$d(t, l) = \min_{\forall k \in K, r \in I} \{d(k, r) | (N_{\min} \leq N_k + N_r \leq N_{\max}) \text{ or } (r \in K, N_k + N_r \leq N_{\max})\}$$

的类别 I_t, I_l , 将两者合并.重复这一步骤直至聚类结果中没有类内样本数过少的类别,即 $\Psi = \emptyset$ 。

在聚类过程中,合并得到新类以后,如果对新类重新训练参数,那么聚类的计算量会相当庞大.我们可以使用某些距离更新方式对新类与其他类之间的距离进行计算,使得聚类过程大大加快.如果进行合并的两类的标

号为 p, q , 合并之后的类别标号为 t , 可用以下方式进行距离的更新:

$$d^*(r, t) = \max(d(r, p), d(r, q)), \quad (4)$$

其中 $d^*(r, t)$ 为更新后的两类之间的距离, $d(r, p), d(r, q)$ 为原来的两类之间的距离.

2 识别确认方法

2.1 帧级别的联合分数

对于连续 HMM, 一般采用高斯混合密度来表示状态输出函数, 如果混合数足够多, 任何一种随机分布函数都可以由高斯分布的线性组合来表示, 所以用混合高斯密度描述 CHMM 中状态的概率输出特性是很合适的.

在搜索过程中, 第 t 帧特征 o_t 在状态 i 时的得分 $\gamma_i(o_t)$ 为输出概率密度 $l_i(o_t)$, 即

$$\gamma_i(o_t) = l_i(o_t) = \sum_{k=1}^M \zeta_{ik} p_{ik}[o_t], \quad \sum_{k=1}^M \zeta_{ik} = 1, \quad (5)$$

其中 ζ_{ik} 是状态 i 的第 k 个混合高斯分布 $p_{ik}[o_t]$ 的权重系数, $p_{ik}[o_t]$ 由下式计算得到:

$$p_{ik}[o_t] = \frac{1}{\sqrt{2\pi} |\Sigma_{ik}|^{1/2}} \exp\left\{-\frac{1}{2}(o_t - M_{ik})^T \Sigma_{ik}^{-1} (o_t - M_{ik})\right\}, \quad (6)$$

M_{ik}, Σ_{ik} 分别为该高斯分布的均值向量和协方差矩阵.

如果在计算得分 $\gamma_i(o_t)$ 时考虑识别置信度, 即将用于确认的参数结合到识别过程中, 求联合得分 $\hat{\gamma}_i(o_t)$, 则搜索得到的最佳路径将是输出概率和置信度分数联合得分最高的识别结果序列. 通过设定适当的权重, 计算出似然得分和置信度得分的综合得分, 即可在关键词的正确检出率和非关键词的正确拒识率之间得到满足系统性能要求的平衡.

在此, 我们设定联合得分 $\hat{\gamma}_i(o_t)$ 由下式所示:

$$\hat{\gamma}_i(o_t) = l_i(o_t)^{w_1} \cdot CM_i(o_t)^{w_2}, \quad (7)$$

其中 $CM_i(o_t)$ 为第 t 帧特征 o_t 在第 i 个状态的置信度分数, w_1, w_2 分别为输出概率和置信度分数的权重. 在实际计算中, 对联合得分 $\hat{\gamma}_i(o_t)$ 取对数:

$$\log \hat{\gamma}_i(o_t) = w_1 \log l_i(o_t) + w_2 \log CM_i(o_t). \quad (8)$$

在式(2)中, 概率似然比是由整段语音分别对关键词模型和各择模型的概率似然值计算得到的. 在帧同步的识别搜索过程中, 我们必须累计每一帧在相应状态上的概率似然比值, 第 t 帧特征 o_t 在状态 i 时的帧级别上的概率似然比值 $LR_i(o_t)$ 由下式给出:

$$LR_i(o_t) = \log l_i(o_t) - \log l_i^a(o_t) = \log p(o_t | \lambda_i) - \log p(o_t | \lambda_i^a), \quad (9)$$

其中, λ_i 为当前模型状态 i 的参数, λ_i^a 为相应的各择模型状态 i 的参数.

概率似然比值 LR 的变化范围很大, LR 得分不能直接累计到搜索路径得分中, 否则会严重影响在搜索过程中的剪枝, 必须对 LR 得分进行适当的预处理.

我们可以对 LR 得分求 $[0-1]$ 间光滑可导的 Sigmoid 函数值:

$$\log CM = \Xi(LR) = \log \frac{1}{1 + \exp(-\alpha \cdot (LR - \beta))}. \quad (10)$$

这里, α 和 β 分别是控制 Sigmoid 函数斜率和位置的常量.

为了减少计算的复杂度, 近似式(11)如下:

$$\Xi(LR) = \begin{cases} \alpha \cdot (LR - \beta), & \text{当 } LR < \beta - 6/\alpha \text{ 时} \\ 0, & \text{当 } LR > \beta + 6/\alpha \text{ 时} \\ \Xi(LR), & \text{其他情况} \end{cases}. \quad (11)$$

2.2 其他级别的联合分数

上面讨论的是在搜索过程中加入帧级别(frame-level)的概率似然比值, 局部的帧级别似然比值可能导

致路径得分发生较大的波动,可以在搜索过程中对其进行适当的平滑,只在音节的结束处加入音节级别(syllable-level)的概率似然比值.假设音节 n 结束于第 t 帧,定义音节级别概率似然比值为

$$SLR_n(o_t) = 1/\tau \cdot \sum_{t-\tau < l \leq t} LR_l(o_l). \quad (12)$$

其中 τ 为音节 n 驻留的时间长度, $LR_l(o_l)$ 是第 l 帧特征在相应的第 i 个状态上帧级别的概率似然比值.同理,可以进行进一步的平滑,在关键词的结束处加入词级别(keyword-level)的概率似然比值.假设词 k 结束于第 t 帧,定义词级别概率似然比值为

$$KLR_k(o_t) = 1/N \cdot \sum_n SLR_n(o_t). \quad (13)$$

3 实验

3.1 实验平台

本文所采用的语音数据库包括两部分:(1) 清华大学智能技术与系统国家重点实验室录制的 CIDS 语音数据库;(2) 专门录制的 10 个人的地名数据库,每个人说了 100 个不同的中文地名,每个词重复 10 遍,其中前 50 个词作为集内词集,后 50 个词作为集外词集,集内词集中前 3 遍作为训练数据,后 7 遍作为实验数据(注:这个实验平台是电梯轮椅控制系统,特定人的,所以只用了 3 个训练数据).我们录制的是 11025Hz 采样率、16 位和单声道的语音信号.

识别系统的识别基元使用无跨越从左向右的半连续 HMM 模型,每个 HMM 状态数大于 3,状态数目依训练数据语音帧数而定,码本空间大小为 64.生成的 MFCC 特征向量是 39 维的向量,分别是 12 维的 MEL 特征系数及其一二阶差分,一维的能量特征系数及其一二阶差分.

3.2 实验结果及讨论

3.2.1 备择模型训练实验及其结果

此实验使用 CIDS 数据库中 60 个男声的数据,每个男声包括 100 个无调拼音的数据,其中 40 个男声是训练用的,20 个男声是测试用的,每个男声前 50 个音节属于集内词集,后 50 个音节是集外词集.所期望的聚类分类数目是 5 类.我们比较本文提出的通过聚类训练备择模型和一般备择模型训练方法两者之间的性能优劣.一般备择模型训练方法是以前 n 个与某关键词最相似的关键词的样本作为训练集,训练该关键词的备择模型.

我们通过集内词和集外词似然比得分的直方图(检出错误平衡图,detection error tradeoff,简称 DET)之间的重叠部分面积的大小来评价确认方法的性能.DET 图中重叠部分较小的确认系统意味着区分集内词和集外词的能力更强.

训练出两种备择模型,方法 1 是采用改进合并分级聚类的方法进行训练,方法 2 是直接使用前 5 个最相似的关键词的样本进行训练.分别计算所有集内词和集外词对两种备择模型的概率似然比,统计出这些词在每个似然比区间段内出现的次数,然后作出相应的 DET 图比较两种备择模型的确认性能.

在图 1 中,聚类训练方法对应的两条曲线之间的重合面积比普通训练方法的要小,即区分集内词和集外词的能力更强;选择合适的确认阈值,聚类训练方法可以获得更高

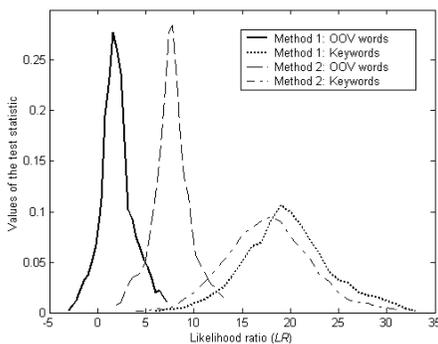


Fig.1 The DET of alternative models

图 1 备择模型确认性能的比较

的正确检出率和拒识率.

3.2.2 识别确认性能

首先我们不加任何确认过程,比较普通的仅累计似然度得分的识别方法和本文提出的联合得分识别确认

方法对集内词的检出性能.我们设定 $w_1=1$,比较联合得分识别确认方法在不同 w_2 值情况下的检出率,结果见表 1,其中当 $w_2=0$ 时,联合得分识别确认方法相当于普通识别方法.

Table 1 The detection performance of combination score method ($w_1=1$)

表 1 联合得分识别确认方法的检出性能($w_1=1$)

w_2	0	0.2	0.4	0.6	0.8	1
Detection rate (%)	97.187	96.246	95.981	91.625	87.652	82.311

实验结果表明,随着 w_2 系数的变大,对集内词的检出率会不断变小,因为模型参数是用最大似然方法训练的,所以在识别过程中加上概率似然比值得分,必然会导致对集内词检出率的下降.在以后的实验中,我们取 $w_2=0.4$.

我们加上初步的确认过程,即在搜索结束以后计算帧平均得分,然后与某一设定的阈值进行比较,以决定该识别结果是被接受或者被拒绝.阈值的变化,可以导致系统检出率和误警率相应发生变化.我们作出操作点特征曲线(receiver operating characteristic,简称 ROC)图,来评价 $w_2=0$ 和 $w_2=0.4$ 时系统的性能.如图 2 所示,在相同的误警率条件下,混合得分识别确认方法的检出率要大大高于普通识别方法的检出率.

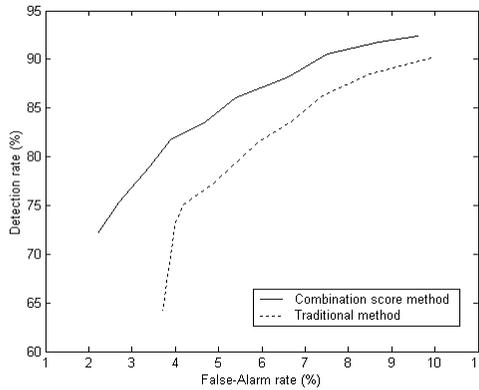


Fig.2 Verification performance comparison of combination score method with traditional recognition method

图 2 联合得分识别确认方法与普通识别方法确认性能的比较

4 结论

本文对语音识别系统中的确认算法进行了深入的研究,结合一阶段和二阶段确认方法的优点,在识别阶段加入概率似然比值确认得分,对传统的 Viterbi 识别进行了一定的改进,提出了联合得分的一阶段识别确认方法.其中备择模型采用基于改进合并分级聚类算法的训练方法,在聚类的过程中,对模型间距离定义、类之间距离更新以及类内样本数目控制进行了探讨,实验结果表明取得了较好的聚类结果,备择模型的确认性能有了大幅度的提高.一系列实验和语音控制系统的实际应用证明,本文提出的一阶段联合得分识别确认方法可以有效地提高系统的确认性能.

References:

[1] López-Cózar R, Rubio AJ, Garcia P, Segura JC. A new word-confidence threshold technique to enhance the performance of spoken dialogue systems. In: Proceedings of the European Conference on Speech Communication and Technology. 1999. 1127~1130.

[2] Wilpon JG, Rabiner LR, Lee C-H, Goldman ER. Automatic recognition of key-words in unconstrained speech using hidden Markov models. IEEE Transactions on Acoustics Speech and Signal Processing, 1990,38(11):1870~1878.

[3] Rose RC. Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition. Computer Speech and Language, 1995,9(4):309~333.

- [4] Sukkar RA, Wilpon JG. A two pass classifier for utterance rejection in keyword spotting. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Minneapolis: Causal Productions Pty Ltd., 1993. 451~454.
- [5] Yan BF, Guo R, Zhu XY, Zhang B. An approach of keyword spotting based on HMM. In: IEEE Proceedings of the 3rd World Congress on Intelligent Control and Automatic. 2000. 2757~2759 (in Chinese with English abstract).
- [6] Tsiporkova T, Vanpoucke F, Van Hamme H. Evaluation of various confidence-based for isolated word rejection. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul: Causal Productions Pty Ltd., 2000. 1819~1822.
- [7] Rahim MG, Lee CH, Juang BH. Discriminative utterance verification for connected digits recognition. IEEE Transactions on Speech and Audio Processing, 1997,5(3):266~277.
- [8] Liu J, Zhu XY. Utterance Verification based on dynamic garbage evaluation approach. Chinese Journal of Computers, 2001,24(5):480~486 (in Chinese with English abstract).
- [9] Boulard H, D'hoore B, Boite J-M. Optimizing recognition and rejection performance in wordspotting systems. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Adelaide: Causal Productions Pty Ltd., 1994. 373~376.
- [10] Wessel F, Macherey K, Schluter R. Using word probabilities as confidence measures. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Seattle: Causal Productions Pty Ltd., 1998. 225~228.
- [11] Hazen TJ, Bazzi I. A comparison and combination of methods for OOV word detection and word confidence scoring. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City: Causal Productions Pty Ltd., 2001. 397~400.
- [12] Ricardi G, Gorin AL, Ljolje A, Riley M. A spoken language system for automated call routing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Munich: Causal Productions Pty Ltd., 1997. 1143~1146.

附中文参考文献:

- [5] 严斌峰,郭锐,朱小燕,张钹.基于隐马尔可夫模型的关键词检出方法.见:第3届世界智能控制和自动化大会论文集.2000. 2757~2759.
- [8] 刘俊,朱小燕.基于动态垃圾评价的语音确认方法.计算机学报,2001,24(5):480~486.