一种联合语种识别的新型大词汇量连续语音识别算法

单煜翔 [1] 邓 妍 [1] 刘 加 [1]

摘 要 提出了一种联合语种识别的新型大词汇量连续语音识别 (Large vocabulary continuous speech recognition, LVCSR) 算法, 并构建了实时处理系统. 该算法能够充分利用语音解码过程中收集的音素识别假设, 在识别语音内容的同时识别语种类别. 该系统可以应用于多语种环境, 不仅可以以更小的系统整体计算开销替代独立的语种识别模块, 更能有效应对在同一段语音中混有非目标语种的情况, 极大地减少由非目标语种引入的无意义识别错误, 避免错误积累对后续识别过程的误导. 为将语音内容识别和语种识别紧密整合在一个统一语音识别解码过程中, 本文提出了三种不同的算法对解码产生的音素格结构进行调整 (重构): 一方面去除语音识别中由发音字典和语言模型引入的特定目标语种偏置, 另一方面在音素格中包含更加丰富的音素识别假设. 实验证明, 音素格重构算法可有效提高联合识别中语种识别的精度. 在汉语为目标语种、汉英混杂的电话对话语音库上测试表明, 本文提出的联合识别算法将集外语种引起的无意义识别错误减少了 91.76 %, 纯汉字识别错误率为 54.98 %.

关键词 语音识别, 语种识别, 集外语种问题, 音素格重构

DOI 10.3724/SP.J.1004.2012.00366

A Novel Large Vocabulary Continuous Speech Recognition Algorithm Combined with Language Recognition

SHAN Yu-Xiang[1] DENG Yan[1] LIU Jia[1]

Abstract In this paper, a novel large vocabulary continuous speech recognition (LVCSR) algorithm combined with language recognition is proposed, and a real-time processing system is developed. This algorithm can make full use of phonetic hypotheses collected during decoding, and identify language types simultaneously. In a multilingual environment, this algorithm can not only take the place of a standalone language recognizer at a lower system overall computational cost, but also effectively cope with the case where target and non-target languages mix in a single utterance. It can significantly reduce speech recognition error introduced by non-target language, and avoid error accumulation which may mislead the subsequent decoding procedure. In order to tightly combine the content and language recognition into a unified decoding procedure, three different phone lattice reconstruction algorithms are also proposed to eliminate pronunciation and grammar restrictions introduced by the target language′s dictionary and language model of the LVCSR decoder, and to encode lattices with richer phonetic information. Experiments show that the lattice reconstruction algorithms can significantly improve language recognition accuracy in the combined recognition. Evaluated on a Mandarin/English mixed conversational telephone speech corpus where Mandarin is the target language, the proposed algorithms reduced the recognition error introduced by non-target language by 91.76 %, and achieved a character error rate of 54.98 %.

Key words Speech recognition, language recognition, out-of-language problem, phone lattice reconstruction

随着越来越多的自动语音识别系统工作于多语种环境, 语种识别技术作为一种必备的分类与过滤手段也越来越受到重视. 在当前主流的应用系统中, 语种识别往往作为一个独立的模块, 位于其他识别或处理模块之前, 或者与其他模块并行工作, 在适当的时候根据语种识别结果对处理流程进行调整[1]. 不论是哪种工作模式, 独立的语种识别模块都引入了不小的额外计算开销, 并且它们都只能对整段的语音进行语种分类, 而对于在同一段语音中混杂有多种语种的情况无能为力. 这种多语种混杂在网络上出现的情况越来越普遍. 通常将待识别语音信号中包含有非指定语种的情况称为语音识别中的集外语种问题 (Out-of-language, OOL). 集外语种会在语音识别结果中引入大量无意义的随机识别错误,

收稿日期 2011-06-03 录用日期 2011-10-08
Manuscript received June 3, 2011; accepted October 8, 2011
高技术研究发展计划 (国家 863 计划) (2008AA02Z414, 2008AA040201), 国家自然科学基金 (60776800, 61005019), 国家自然科学基金委员会与香港研究资助局联合科研基金 (60931160443) 资助
Supported by High Technology Research and Development Program of China (863 Program) (2008AA02Z414, 2008AA040201), National Natural Science Foundation of China (60776800, 61005019), National Natural Science Foundation of China and Research Grants Council of Hong Kong (60931160443)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 清华大学电子工程系清华信息科学与技术国家实验室 北京 100084
1. Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084

以这些随机错误作为历史路径, 语言模型就会错误地引导后续解码过程, 干扰集内语种的识别. 集外语种问题在语音检索系统应用中经常出现, 逐渐受到研究人员的重视. 文献 [2−4] 提出了一种在关键词检测系统中, 采用通用置信度过滤集外语种的方法. 尽管该方法运算简单, 并能有效地减少由集外语种造成的关键词虚警, 但从本质上讲, 它仅仅是另一种关键词置信度的计算方法, 无法提供语种类别信息, 因此也不可能替代独立的语种识别模块. 鉴于目前以音素识别为前端的语种识别技术已相当成熟, 具有很好的识别效果. 如果将其与语音识别技术更为紧密地融合, 必能更有效地解决集外语种问题, 弥补独立语种识别模块的缺陷.

基于以上思想, 本文提出了一种联合语种识别的新型大词汇量连续语音识别 (Large vocabulary continuous speech recognition, LVCSR) 算法. 该算法将 LVCSR 和基于音素识别－向量空间建模 (Phoneme recognition-vector space modeling, PRVSM)[5] 的语种识别相融合, 能够在识别语音内容的同时对语种类别进行实时检测, 可有效地分离任何形式的集外语种, 完全替代独立的语种识别模块. 由于该算法充分复用了语音识别过程中收集的音素识别假设, 节省了对语音进行二次特征提取和解码的时间, 所以基于该算法构建的识别系统, 其整体时间效率优于采用独立语种识别和语音识别模块的传统系统.

PRVSM 语种识别技术通过对音素之间的统计关系 (n 元文法) 建模来区分不同的语种. 其前端一般采用语种无关的音素识别器将语音转化为音素序列或音素格[6], 后端在音素序列或格上进行 n 元文法统计建模. 但在 LVCSR 中, 发音字典限制了词内音素的连接关系, 语言模型约束了词间音素的连接关系, 使得音素识别结果严重偏向于特定目标语种. 为了在联合识别中克服这一目标语种偏置, 本文提出了三种不同的音素格重构算法, 一方面打破字典和语言模型的约束, 另一方面为语种识别提供更加丰富的音素识别假设. 实验证明, 即使在语音长度很短的情况下, 这些算法仍能有效地提高语种识别的精度.

除了基于音素识别的方法, 语种识别也可以利用声学层特征[7]. 这类系统通常采用与语音识别不同的长时相关特征 (如偏移差分倒谱[8]), 算法与语音识别差距较大, 不易复用语音识别解码已产生的信息, 因此在本文中没有采用.

本文的组织结构如下: 首先介绍语音内容和语种联合识别的基本假设和算法, 接下来详述本文提出的三种音素格重构方法, 并简要说明采用的 VSM 语种分类器, 最后是实验结果与分析.

1　语音内容与语种联合识别算法

语音识别的任务是将语音转化为文字. 给定输入特征序列 O, 传统的语音识别算法在由发音字典、声学模型和语言模型等知识源组成的搜索空间中, 通过 Viterbi 算法寻找出最佳词串 W^*, 满足:

$$W^* = \arg\max_{W} p\left(W|O\right) \tag{1}$$

若要对语音内容和语种进行联合识别, 则需要在式 (1) 中添加语种相关项, 修改后的目标函数为

$$\begin{aligned}(W^*, L^*) = \arg\max_{W,L} p\left(W, L|O\right) = \\ \arg\max_{W,L} \left\{p\left(L|W, O\right) p\left(W|O\right)\right\}\end{aligned} \tag{2}$$

其中, L^* 表示最优的语种序列. 如果采用 PRVSM 语种识别, 则可以重用语音识别产生的结果, 式 (2) 可以进一步写为

$$(W^*, L^*) = \arg\max_{W,L} \left\{\sum_{P} p\left(L|P\right) p\left(P|W\right) p\left(W|O\right)\right\} \tag{3}$$

其中, P 表示上下文无关的音素序列; $p\left(W|O\right)$ 为语音识别项, 可采用与 LVCSR 相同的 Viterbi 算法求解; $p\left(L|P\right)$ 为语种识别项, 本文采用 VSM 方法求解[5]; $p\left(P|W\right)$ 表示词串到音素串的映射关系, 如前所述, 受限于发音字典和语言模型, 从而影响语种识别的性能 (见第 4 节实验). 为解决这一问题, 本文提出采用音素格重构算法改善该映射关系, 抑制发音字典和语言模型的影响, 生成适合语种识别的音素格. 由于直接对式 (3) 进行优化非常困难, 并且没有必要在每一帧都做语种识别 (因为语种反映的是语音的长时特征), 所以综合上面讨论的三个部分, 将语音内容和语种联合识别算法步骤概括如下:

步骤 1. 按照与传统 LVCSR 相同的方式对输入语音进行 Viterbi 解码, 并收集音素识别假设;

步骤 2. 在第 n 个可能的语种转换点 t_n, 根据已收集的音素识别假设在时间区间 $[t_{n-1}, t_n]$ 上重构音素格;

步骤 3. 利用重构后的音素格进行语种识别, 并记录语种识别结果 (可根据需要记录多候选结果);

步骤 4. 若语种识别结果为目标语种, 则返回步骤 1 继续解码过程; 若语种识别结果并非目标语种, 则重置解码器的语言模型状态后, 再返回步骤 1 继续解码 (已解码的识别结果作为历史路径保留);

步骤 5. 语音处理完成后, 对语音识别和语种识别结果进行回溯, 得到最佳词串 W^* 和最佳语种序列 L^*.

语种转换点的选择有多种方式. 本文选择在语音的自然停顿点, 并且对两次相邻停顿点的最小时间间隔进行约束 (即选取 $T_{\min}$ 作为门限, 使得 $t_n - t_{n-1} \le T_{\min}$, $t_0 = 0$), 以保证语种识别和语音识别结果的可靠性. 停顿点的检测可以利用静音模型与解码同步进行, $T_{\min}$ 的选择通过实验确定 (见实验部分).

2 音素格重构算法

在 LVCSR 中, 音素识别假设的收集和音素格的生成通常有两种方式:

1) 词同步对齐方式: 这是 LVCSR 中最常见的音素格生成方式. 该方式中, 每当一个词识别假设被添加进入历史路径, 为得到该词识别假设而经历的音素级路径也被随之保留; 当词识别假设被剪枝, 与之相关的音素识别假设也随之删除. 采用这种方式, 最终可以得到与词格严格对应的音素格, 本文称之为“原始音素格”.

2) 音素集合方式: 该方式中, 不考虑音素对应的词级信息. 所有的音素识别假设一旦生成, 就被添加到同一个音素集合中去. 该集合中的每一个成员都记录音素的标识、出现时间以及似然分数. 若在相同的位置识别出了相同的音素, 则拥有较高似然分数的识别假设被保留. 一旦音素识别假设被添加到集合中, 剪枝便与之毫无关系. 即使音素所隶属于的词被剪掉, 该音素依然会保留在集合中. 因此, 利用音素集合可以重构出信息更为丰富而又不受限于发音字典和语言模型的音素格.

音素格重构算法主要从两个方面改善联合语种识别的性能. 首先, 采用启发式算法对原始音素格中识别假设的时间点和边进行聚类、对准, 打破发音字典中固有的音素间连接关系和语言模型所约束的词间连接关系; 其次, 采用 “音素集合方式”, 在解码过程中收集更加丰富的音素识别假设, 添加到对准后的音素格中作为语种分类器的输入. 尽管重构后的音素格, 时间点经过了聚类和对准, 变得不再准确, 但是音素之间相互连接的顺序关系保持不变, 因此对语种识别没有影响.

为描述方便, 首先定义如下符号表示. 音素格 $L = (N, A, n_{\text{start}}, n_{\text{end}})$ 表示一个有向无环图, 其中 N 和 A 分别表示节点和边的集合, n_{start} 和 n_{end} 分别表示格的开始和结束节点. $\forall n \in N$, $T[n]$ 表示其所对应的时间点. 相应地, $N[t] \in N$ 表示时间点 t 所对应的节点. $\forall a \in A$, $S[a]$ 和 $E[a]$ 分别表示它的起始与终止节点, $I[a]$ 表示它所对应的识别假设, $ac[a]$ 和 $pr[a]$ 分别表示它的声学模型分数和后验概率. 对于一个音素识别假设 p, $S[p]$ 和 $E[p]$ 分别表示它出现的起始与终止时间, $I[p]$ 表示它的音素标识, $ac[p]$ 表示它的声学分数.

2.1 原始音素格时间点对齐

本文提出的第一种音素格重构算法采用与混淆网络[9] 相类似的节点、边聚类过程, 将原本非线性的音素格转化为一个线性图, 并维持原始格中的序关系不变. 给定一个音素格 L, 可以定义边的等价类与偏序关系: 对于 $a \in A$, 等价类 $[a]$ 表示所有与 a 对准的边的集合. 对于 $a, b \in A$, 称 $a \le b$ 当且仅当 $a = b$ 或 $E[a] = S[b]$ 或 $\exists c \in A$ 使得 $a \le c \le b$. 如果 $\forall a_1 \in [a]$, $b_1 \in [b]$, 都有 $a_1 \le b_1$, 那么称 $[a] \le [b]$. 时间点对齐算法的本质即是将原始音素格划分为若干等价类, 并保证各个等价类之间的偏序关系与原始音素格一致. 算法如下:

步骤 1. 对输入语音特征进行解码, 按照词同步对齐方式收集音素识别假设, 生成原始音素格.

步骤 2. 利用原始音素格中的所有音素识别假设 p 和它们的起始与终止时间 t_1、t_2 初始化等价类:

$$C_{p,t_1,t_2} = \{a : I[a] = p, S[a] = t_1, E[a] = t_2, a \in A\} \tag{4}$$

令 $X = \{C_{p,t_1,t_2} : \forall p, t_1, t_2\}$ 为所有初始等价类集合.

步骤 3. 从 X 中找出两个最相似, 但又无序的等价类 C_1^* 和 C_2^*, 合并为一个新的等价类 $C_{\text{new}} = C_1^* \cup C_2^*$:

$$(C_1^*, C_2^*) = \arg \max_{\substack{C_1 \in X, C_2 \in X \\ C_1 \not\lesssim C_2, C_2 \not\lesssim C_1}} \text{SIM}(C_1, C_2) \tag{5}$$

两个等价类之间的相似度定义为

$$\text{SIM}(C_1, C_2) = \max_{\substack{a_1 \in C_1 \\ a_2 \in C_2}} \text{overlap}(a_1, a_2) \cdot pr[a_1] \cdot pr[a_2] \tag{6}$$

式中 $\text{overlap}(a_1, a_2)$ 表示两条边的时间重合度, 定义为两条边重合的时间长度除以它们各自长度之和.

步骤 4. $\forall C \in X$, 更新 X 中等价类之间的偏序关系: 若 $C \le C_1^*$ 或 $C \le C_2^*$, 则 $C \le C_{\text{new}}$; 若 $C_1^* \le C$ 或 $C_2^* \le C$, 则 $C_{\text{new}} \le C$.

步骤 5. $\forall (C_1, C_2) \in X \times X$, 更新 X 中等价类之间的偏序关系: 若 $C_1 \le C_1^*$ 且 $C_2^* \le C_2$, 或者 $C_1 \le C_2^*$ 且 $C_1^* \le C_2$, 则 $C_1 \le C_2$.

步骤 6. 令 $X = X \cup \{C_{\text{new}}\} \setminus \{C_1^*, C_2^*\}$.

步骤 7. 重复步骤 3 ∼ 步骤 6, 直到 X 中再没有无序等价类.

步骤 8. 输出时间点对齐后的音素格.

该时间点对齐算法的时间复杂度是 $\mathrm{O}(|A|^3)$. 该算法与标准的混淆网络生成算法的区别在于计算等价类相似度时只考虑两类的时间重合度, 而不计入音素发音的相似度或混淆度, 因为采用 VSM 方法进行语种建模时主要关心的是音素的上下文关系, 而非音素之间的相似或易混程度. 对该聚类算法正确性和收敛性的证明与标准的混淆网络生成算法相同[9], 不再赘述.

2.2 最优路径时间点对齐

第 2.1 节利用时间点对齐算法得到了对原始语音的一个时间划分. 事实上, 语音识别生成的最优候选路径本身就提供了这样一种划分. 本文提出的第二种音素格重构算法放弃了对原始音素格的依赖, 首先采用最优音素级候选路径作为时间划分, 将原始语音分割为若干个时间段, 然后从音素集合中为每一时间段选取最有可能出现的 n 个音素识别假设, 组成重构后的音素格. 算法描述如下:

步骤 1. 对输入语音特征进行解码, 按照词同步对齐方式和音素集合方式收集音素识别假设, 生成最优音素候选路径和音素集合;

步骤 2. 从最优音素候选路径获得初始的时间段划分 $B = \{t_0, t_1, \cdots, t_M\}$, 其中 M 为音素边界个数;

步骤 3. 初始化音素格 L: $N = \{N[t] : \forall t \in B\}$, $A = \varnothing$, $n_{\mathrm{start}} = N[t_0]$, $n_{\mathrm{end}} = N[t_M]$;

步骤 4. $\forall i = 1, \cdots, M$, 从音素集合中获得具有起始时间 t_{i-1}, 终止时间 t_i, 声学分数最高的 n 个音素识别假设, 存放在列表 P 中;

步骤 5. 对 P 中的每个音素识别假设 p, 建立一条新的边 a: $S[a] = S[p]$, $E[a] = E[p]$, $I[a] = I[p]$, $ac[a] = ac[p]$;

步骤 6. 输出生成的音素格 L.

假设音素集合中对应 B 的每个时间段划分的平均音素个数为 k. 由于实际中使用的 n 通常非常小, 所以该算法的时间复杂度近似为 $\mathrm{O}(M \cdot k)$.

2.3 简单音素识别假设拼接

第 2.2 节采用的最优候选路径仍然是在发音字典和语言模型的指导下产生, 或多或少会受到它们的影响. 为了进一步摆脱其约束, 本文提出的第三种音素格重构算法完全抛弃了时间点对齐, 转而尝试从音素集合中重建音素格. 算法描述如下:

步骤 1. 对输入语音特征进行解码, 仅按照音素集合方式收集音素识别假设, 得到音素集合;

步骤 2. 初始化音素格 L: $N = \{N[t] : t = 0, 1, \cdots, M\}$, $A = \varnothing$, $n_{\mathrm{start}} = N[0]$, $n_{\mathrm{end}} = N[M]$, M 为语音帧数;

步骤 3. $\forall i = 1, \cdots, M$, 从音素集合中获得具有起始时间 $i-1$, 终止时间 i, 归一化声学分数最高的 k 个音素识别假设, 存放在列表 P 中. 归一化声学分数定义:

$$score_{\mathrm{Norm}}(p) = \frac{\log ac[p]}{E[p] - S[p]} \tag{7}$$

步骤 4. 对 P 中的每个音素识别假设 p, 建立一条新的边 a: $S[a] = S[p]$, $E[a] = E[p]$, $I[a] = I[p]$, $ac[a] = ac[p]$;

步骤 5. 移除 L 中不可达的节点和边, 并输出 L.

步骤 5 是必需的, 因为音素集合中存在大量来自于被剪枝路径的音素识别假设. 在多数情况下, 它们能够与未被剪掉的路径相连通, 但也常常会存在不可达的情况 (从 n_{start} 不可达或从 n_{end} 不可达), 这会对语种识别时做 n 元文法统计造成影响.

3 VSM 语种识别后端

利用重构后的音素格, 我们采用 VSM 方法进行语种识别, 从而将语种相关信息引入到 LVCSR 过程中, 实现语音内容和语种类别的联合识别. VSM 方法的出发点是: 利用一个高维的超向量对某个音素序列中所有可能出现的音素及其组合 (统称 n 元文法) 进行描述. 由于在不同语种的音素序列中, n 元文法出现的概率是不同的, 所以可以据此区别语种类别. 目前, 主流的 VSM 语种识别方法都是在音素格上进行 n 元文法统计, 本文采用的方法与文献 [5] 相似.

给定一个音素格 L, 假设 $S = s_1, \cdots, s_N$ 表示 L 中一条可能的音素序列, 则 L 的 n 元文法统计量计算方法如下:

$$\begin{aligned}\mathrm{count}(s_i, \hat{s}_i|L) = E_S[\mathrm{count}(s_i, \hat{s}_i|S)] = \\ \sum_{S \in L} p(S|L)\,\mathrm{count}(s_i, \hat{s}_i|S)\end{aligned} \tag{8}$$

其中, $\hat{s}_i = s_{i-(n-1)}, \cdots, s_{i-1}$ 表示 n 元文法的历史信息, n 为阶数. count 函数返回在音素序列 S 中 n 元文法 $(s_i, \hat{s}_i)$ 出现的次数. $p(S|L)$ 表示音素序列 S 在音素格 L 出现的概率, 可通过标准的前后向算法[6] 计算得到. 通过 n 元文法统计量, 可以根据式 (9) 计算 n 元文法 $(s_i, \hat{s}_i)$ 在音素格 L 中出现的概

率:

$$p\left(s_i,\hat{s}_i|L\right)=\frac{\mathrm{count}\left(s_i,\hat{s}_i|L\right)}{\sum\limits_j \mathrm{count}\left(s_j,\hat{s}_j|L\right)} \tag{9}$$

假设某语种的音素集包含 N 个不同的音素, 也就对应了 $D=\sum_{i=1}^{n}N^i$ 个不同的 n 元文法. 给定一个音素格 L, 将式 (9) 计算得到的 n 元文法概率映射到一个 D 维的稀疏超向量 $\boldsymbol{V}$, $\boldsymbol{V}$ 中的每一个元素对应该语种的一个可能的 n 元文法在音素格 L 中出现的概率. 由于 $\boldsymbol{V}$ 是一个超高维向量, 一般采用支持向量机 (Support vector machine, SVM) 对向量中各分量的权重进行学习, 由权重向量所表示的超平面来对不同的语种进行分类. 相关研究[10] 表明, 对超向量 $\boldsymbol{V}$ 进行适当的归一化可有效改善识别性能. 因此, 本文采用词频对数似然比 (Term frequency log likelihood ratio, TFLLR) 方法[10] 对 $\boldsymbol{V}$ 进行归一化, 并在 SVM 中选用线性核作为不同超向量之间相似性的度量. 综合这两步的核函数如下:

$$K(L_1,L_2)=\sum_{i=1}^{N}\frac{p\left(s_i,\hat{s}_i|L_1\right)}{\sqrt{p\left(s_i,\hat{s}_i|\mathrm{all}\right)}}\cdot\frac{p\left(s_i,\hat{s}_i|L_2\right)}{\sqrt{p\left(s_i,\hat{s}_i|\mathrm{all}\right)}} \tag{10}$$

式中, L_1 和 L_2 表示两个输入的音素格, $p(s_i,\hat{s}_i|\mathrm{all})$ 在训练集中的所有音素格上计算得到.

4 实验及结果分析

4.1 实验设置

本节在 CallHome 数据库的开发集和测试集的中、英文电话对话语音数据上验证本文所提出的语音内容和语种联合识别算法的性能. 实验分为两个部分, 首先以一个简单的汉、英语种分类任务对比不同的音素格重构算法对语种识别性能的影响, 然后选定一种重构算法, 检验联合识别算法在汉、英混杂环境下的识别性能.

实验中, 选用一个 68 k 词的汉语 LVCSR 引擎作为基线系统和联合识别前端. 该识别引擎基于令牌传递算法[11] 实现, 声学模型采用 6 000 状态 32 高斯的跨词三音子模型, 语言模型采用 Trigram. 声学模型训练采用最小音素错误 (Minimum phoneme error, MPE) 准则[12], 训练数据包含从 HKUST、CallFriend、CallHome 训练集和 863 中文数据库中选择的约 300 h 数据. 语言模型训练数据则采用 Gigaword 中文数据库和训练语音标注. 对于语种分类后端, 采用第 3 节中描述的方法构建 VSM 分类器. 训练数据采用选自 CallFriend 和 CallHome 训练集中的约 9 000 段汉、英语音. 每段语音长约 30 s, 均采用语音端点检测算法从原始语音中自动切分得到. 这些语音经过与测试时相同 LVCSR 和音素格重构, 重构后的音素格参与 VSM 训练. 我们的汉语音素集共包含 96 个不同的音素 (不包含静音模型 Sil 和 Sp). 实验中统计了三元文法概率, 得到的超向量维数为 894 048 维.

此外, 为了进一步对比音素格重构算法的性能, 我们还搭建了一个经典的 PRVSM 语种识别系统. 该识别系统采用与前面相同的 VSM 后端, 但采用一个简单的语种无关的音素识别器作为前端来得到音素格.

4.2 音素格重构算法性能比较与 $\boldsymbol{T}_{\min}$ 选择

图 1 以 DET (Detection error tradeoff) 曲线的形式给出了采用不同的音素格重构算法时的语种识别性能, 表 1 的最后一行 (总计) 则给出了相应的等错误率 (Equal error rate, EER). 测试段采用 CallHome 开发集和测试集标注手工切分得到, 并且丢弃了长度小于 0.5 s 的语音. 在汉、英语种分类中存在两种错误: 虚警错误 (把英语语音误判为汉语) 和漏报错误 (把汉语语音误判为英语). DET 曲线反映了在不同工作点下这两种错误的消长情况. 从图中可以看出, 原始音素格性能最差, 其原因是发音字典和语言模型约束了音素间的连接关系, 从而影响了靠这种连接关系进行语种识别的效果; 相比而言, 本文提出的三种音素格重构算法都可以有效地改善以 LVCSR 作为前端时语种识别的性能, 其中以简单音素拼接算法整体性能最优. 这是因为在这三种方法中, 简单音素拼接算法对发音字典和语言模型的约束的去除最为彻底.

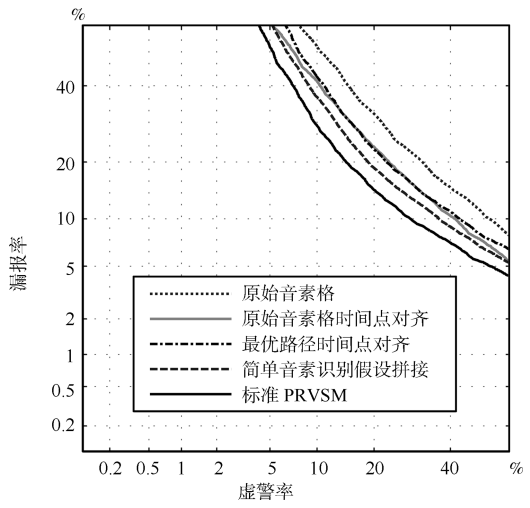


图 1 采用不同音素格重构算法时的语种识别性能

Fig. 1 Language recognition performances with different lattice reconstruction algorithms

为了选择合适的最小相邻语种转换时间门限 $T_{\min}$, 表 1 进一步给出了测试集中的语音段长度分布, 以及在不同长度的语音下, 基于不同的音素格重构算法进行语种识别时的等错误率. 可以看出, 对于各种长度的语音, 尤其是长度大于 3 s 的语音, 采用简单音素识别假设拼接的方法, 都可以获得与 PRVSM 系统相当的性能. 但由于整个测试集中短语音 (长度小于 3 s 的语音) 占据了大多数, 所以在整个测试集上统计得到的性能也偏重于短语音, 这也是图 1 中简单音素识别假设拼接和标准 PRVSM 之间性能差异显得较大的原因.

由于 $T_{\min}$ 控制了联合识别算法中触发语种识别的频度, 考虑到 $T_{\min}$ 过小时语种识别性能欠佳, 并且会降低语音内容识别中语言模型对解码结果的约束作用, 而 $T_{\min}$ 太大又会降低系统响应速度, 所以在后续实验中, 我们选择 $T_{\min} = 3$ s.

表 1　测试集中语音长度分布, 以及在不同语音长度下的音素格重构算法性能对比

Table 1　Test utterance length distribution, and language recognition performance with different utterance lengths and different lattice reconstruction algorithms

测试段分布		等错误率 EER (%)				
长度 (s)	个数	A0	A1	A2	A3	PRVSM
<1	3 549	34.90	32.74	32.48	31.53	29.67
1 ∼ 2	4 114	27.45	23.99	25.15	21.63	19.88
2 ∼ 3	2 651	24.25	18.90	17.52	13.40	12.61
3 ∼ 4	1 674	19.30	15.53	12.66	10.07	8.98
4 ∼ 5	825	17.02	11.34	7.39	5.06	5.05
5 ∼ 6	425	15.76	10.73	7.03	4.17	4.74
6 ∼ 7	228	11.52	6.09	6.08	3.04	2.16
7 ∼ 8	120	11.30	7.29	3.31	2.34	4.92
<8	738	5.67	3.37	2.13	1.47	1.02
总计	14 324	25.00	21.69	21.43	19.26	17.24

注: A0−原始音素格, A1−原始音素格时间点对齐, A2−最优路径时间点对齐, A3−简单音素识别假设拼接, PRVSM−标准 PRVSM 系统

4.3　音素格重构算法时间效率比较

事实上, 简单音素识别假设拼接的方法不仅能够提供良好的识别精度, 还具有最快的音素格重构速度, 从而能够实现更小的系统整体计算开销. 通过实验, 我们发现三种音素格重构算法最耗时的部分都在音素识别假设收集, 而后生成音素格的时间几乎可以忽略不计. 采用词同步对齐方式收集音素识别假设需要维持词串与音素串的严格对应关系, 因此对整个解码过程的影响要大于音素集合方式. 以传统 LVCSR 解码的时间为单位 1 (仅识别字词, 不生成音素格, 不做语种识别), 表 2 给出了采用不同的音素格重构算法做语种识别时的系统整体计算时间对比. 可见, 采用简单音素识别假设拼接方法时, 用于语种识别的时间减少到独立语种识别模块 (标准 PRVSM) 的约 38 %. 这一时间统计采用与第 4.2 节相同的测试集数据, 在一台 Intel Core2 Duo 2.26 G CPU, 3 G 内存的计算机上测得, 测试时仅采用了一个 CPU 核. 正因为简单音素识别假设拼接方法在速度与精度方面的优势, 所以后续实验均采用此方法重构音素格.

4.4　语音内容和语种联合识别性能

本节给出了本文提出的语音内容和语种联合识别算法的性能, 以及与传统 LVCSR 算法的对比. 性能指标采用字错误率 (Character error rate, CER) 以及替代 (Sub)、删除 (Del)、插入 (Ins) 三种错误率和字正确率 (Corr). 测试采用的汉英混杂数据采用如下方式生成: 将 CallHome 测试集和开发集的汉语和英语数据按标注人工切分, 选取其中长度大于 3 s 的语音按照 “汉英汉”、“英汉英”、“汉英”、“英汉”、“汉”、“英” 六种方式随机拼接. 其中, 汉语是待识别的目标语种, 英语是需要过滤的集外语种. 为了对比结果, 将汉语测试段的标注也进行相应的拼接, 在随后的实验中直接对比拼接后大段语音的识别结果. 语种识别的工作点简单选择在等错点.

表 2　采用不同的音素格重构算法做语种识别时的系统整体计算开销 (以传统 LVCSR 为单位 1)

Table 2　The overall computational costs of doing language recognition using different lattice reconstruction algorithms (considering traditional LVCSR as 1)

音素格重构算法	系统整体计算时间
原始音素格	1.14
原始音素格时间点对齐	1.15
最优路径时间点对齐	1.18
简单音素识别假设拼接	1.08
标准 PRVSM	1.21

给定一段待识别的语音, 不论是传统 LVCSR 算法还是本文提出的联合识别算法, 都会对所有的

语音帧进行解码, 从而得到最优的识别结果. 二者的区别在于, 本文提出的联合识别算法在得到最优识别结果的同时, 还能够为每一个语音片段 (即语音中两相邻自然停顿点间的部分) 标出其可能的语种类别并给出置信度. 传统 LVCSR 算法无此功能, 仅能默认所有的语音都属于相同的目标语种, 而无法将集外语种滤除.

首先, 我们在拼接后的测试集上验证传统 LVCSR 的性能. 由于没有集外语种检测, 解码器错误地将英文语音当作中文进行解码, 所以产生了大量无意义的识别结果, 插入错误高达 40.43 % (表 3 第 1 行). 如果手工将识别结果中与英文语音对应的部分去除, 得到的纯汉语语音段识别性能列于表 4 的第 1 行. 可以看出, 在不知道语种分段信息的情况下, 错误地将集外语种 (英语) 语音段的识别结果包含在最终的识别结果中去, 会引入大量无意义的识别结果, 导致插入和替代错误急剧上升, 但这些随机的识别结果也可能 "蒙" 对一些字, 所以删除错误较之有语种分段时要低一些. 总的来说, 集外语种的存在使得字错误率上升了 38.35 %.

表 3 传统 LVCSR 算法与联合识别算法性能比较 (%)

Table 3 Performance comparison between traditional LVCSR algorithm and joint recognition algorithm (%)

算法	CER	Sub	Del	Ins	Corr
传统 LVCSR	90.17	38.56	11.18	40.43	50.26
联合识别	54.98	30.66	22.21	2.12	47.13

表 4 采用人工语种分段时的两种算法性能比较 (%)

Table 4 Performance comparison between the two algorithms using manual language segmentation (%)

算法	CER	Sub	Del	Ins	Corr
传统 LVCSR	51.82	34.84	14.87	2.11	50.29
联合识别	51.70	35.07	14.59	2.04	50.34

对于本文提出的联合识别算法, 利用识别过程中自动生成的语种分段信息, 我们可以将识别结果中集外语种片段的部分去除, 得到的识别性能列于表 3 的第 2 行. 从该结果中可以看出, 在滤除了绝大部分英文语音的错误识别结果后, 插入错误大大降低. 然而, 由于语种识别存在漏报错误, 将部分汉语语音误判为英语, 所以删除错误增加了约一倍. 有趣的是, 尽管语种识别还存在虚警错误, 即将部分英语语音误判为汉语的情况, 但插入错误并没有明显上升, 替代错误反而明显下降 (与表 4 第 1 行相比). 这是因为绝大多数被误判为英语的汉语语音, 其语音识别结果也是不正确的 (经统计, 这些语音的字识别正确率仅为 29.31 %, 替代错误 41.68 %, 删除错误 29.00 %, 插入错误 7.79 %, 字错误率总计为 78.48 %), 所以在统计识别率的过程中, 被错误识别的英语语音段替代了被错误识别的汉语语音段的位置, 使得替代和插入错误没有相应地增加. 这也从另一个角度说明, 语种识别的分数对于语音识别结果来说是一个良好的置信度分数. 采用本文提出的联合识别算法后, 相对于纯汉语语音识别的性能 (表 4 的第 1 行) 字错误率仅上升了 3.16 %, 将由集外语种引入的语音识别错误相对减少了约 91.76 %.

表 4 第 2 行同样给出了采用人工语种分段标注, 即假设语种识别全部正确时的语音识别结果. 该结果可以看作含有独立语种识别前端的系统字识别正确率的上限 (假设通过一些技术手段, 使得独立的语种识别前端能够确定语种转换点, 并据此进行语种识别). 可见, 即使屏蔽语种识别错误, 本文提出的联合识别算法性能仍略优于传统算法. 这说明通过自动地语种识别、分段, 适时地重置语言模型状态, 可以减少集外语种在解码过程中在语言层造成的错误引导和干扰, 避免对后续解码过程的误导.

图 2 进一步给出了联合识别算法在不同的语种识别工作点 (判决门限) 下的字错误率. 图中判决门限为 0 的点对应于等错误点. 从图中可以看出, 随判决门限降低, 虚警率越来越大, 所以插入错误也越来越高; 随门限升高, 漏报率越来越大, 删除错误的比重越来越大.

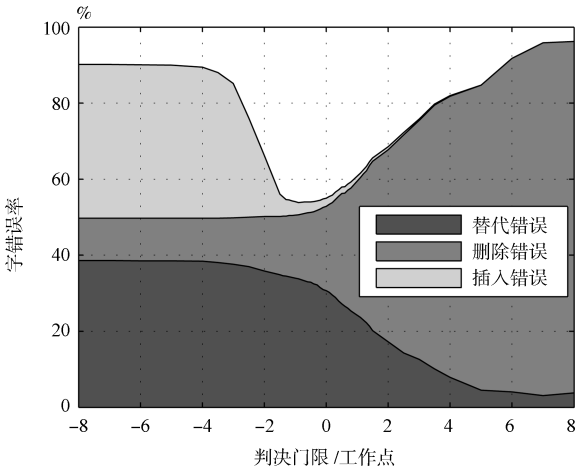


图 2 联合识别算法在不同语种识别工作点下的字错误率

Fig. 2 Character error rate of joint recognition algorithm under different language recognition operating points

4.5 对实验结果的进一步讨论

在本文提出的联合识别算法中, 由于会在可能的语种转换点重置语言模型状态, 所以一定程度上降低了语言模型对解码过程的约束作用. 然而, 从以上实验结果可以看出, 这并没有对语音识别的准确性造成大的影响. 究其原因, 是因为本文提出的联合识别算法具有如下特点:

首先, 音素格重构不干扰正常的 LVCSR 解码. 音素格重构仅仅复用在 LVCSR 解码过程中产生的音素识别假设来生成适用于 VSM 后端的音素格, 而不参与 LVCSR 解码过程中的令牌传递、路径合并、结果回溯、词格生成等过程, 所以不影响语言模型在解码中发挥其应有作用.

其次, 语言模型状态仅在语种识别结果为非目标语种时才需要重置. 如第 1 节算法步骤 4 所述: 当语种识别结果为目标语种时, LVCSR 解码过程将不受任何影响地继续进行, 从而也延续了语言模型的状态和对解码过程约束作用; 当语种识别结果为非目标语种时, 理所应当将语言模型状态重置归零, 以避免错误识别结果对后续解码过程的影响. 即使语种识别中发生了误判, 从第 4.4 节的实验中可知, 语种识别错误的语音片段其 LVCSR 识别结果往往具有极高的错误率, 所以重置语言模型状态也不会带来明显的性能损失.

最后, 语言模型的约束作用需要被解码的语音具有一定长度才能有效地发挥, 这也是我们在实验中选择 $T_{\min} = 3\,\mathrm{s}$ 的另一个原因: 确保被识别的语音段至少包含 10～15 个汉字 (大约一句话的长度), 使得在可能的语种转换点到来之前, 语言模型能够有效地发挥约束作用. 由于语种识别也是基于长时相关特征进行分类, 所以要获得准确、可靠的识别结果, 也需要被识别语音具有一定的长度. 所以, 选择合适的 $T_{\min}$ 对两次可能的语种转换点的时间间隔进行约束, 不仅是保证语种识别可靠性的需要, 也是保证语音识别结果可靠性的需要. 另外, $T_{\min}$ 的选择也依赖于语言模型的阶数. 理论上讲, 语言模型阶数越高, 语音识别结果越准确, 但相应的语言模型约束持续时间也越长, 对语种识别性能潜在的影响也就越大. 本文选用语音识别精度和语言模型约束长度相对均衡的三元文法模型进行实验, 在以后的工作中, 我们将对该问题进行更深入的研究探索.

在第 4.4 节中, 我们选择了长度大于 3 s 的语音进行拼接实验, 构造了一种语种切换不是很频繁的应用场景 (语种切换时间间隔在几秒钟的量级). 而在其他应用场景中, 有可能出现单一语种语音长度小于 $T_{\min}$ 的情况. 如果该单一语种语音长度略小于 $T_{\min}$, 如 2 s～3 s 的情况 (仍以 $T_{\min} = 3\,\mathrm{s}$ 为例), 通过略微放松 $T_{\min}$ 的约束, 本文提出的联合识别算法仍然可以有效工作, 当然语音识别的字错误率也会因此有所上升. 对于更为极端的情况, 例如在大段汉语语音中仅夹杂有一两个英文单词的情况, 则仅采用本文的算法很难准确地定位出非目标语种语音段的位置并将其滤除. 这种情况下, 采用以词为检测单元, 类似于集外词 (Out-of-vocabulary, OOV) 检测的置信度方法[2−4] 更为适合. 但是, 由于检测单元小, 可利用的信息有限, 并且 OOV 检测往往采用语种无关的置信度进行, 所以检测精度不高. 一种可能的改进方法是将本文提出的语种相关的联合识别算法和语种无关的置信度方法相融合, 从而改善算法在短语音情况下的识别性能. 这也是本文下一步工作的另一个研究方向.

5 结论

针对大词汇量连续语音识别中的集外语种问题, 尤其是在同一段语音中出现的集外语种问题, 本文提出了一种联合语种识别的新型大词汇量连续语音识别算法. 该算法以解码得到音素识别假设为特征, 采用向量空间建模技术同步地进行语种识别. 为了克服由发音字典和语言模型引入的目标语种偏置, 提出了三种音素格重构算法, 一方面打破发音字典和语言模型的约束, 另一方面为语种识别提供更加丰富的音素信息. 实验证明, 基于简单音素识别假设拼接的音素格重构算法可以提供与独立的语种分类器相似的性能, 而本文提出的联合识别算法在多语种混杂情况下可以极大地减少由集外语种引入的替代错误和插入错误. 此外, 由于充分复用了解码过程中产生的信息, 所以基于该算法构建的语音识别系统其整体时间效率优于采用独立语种识别和语音识别模块的传统系统, 用于语种识别的时间减少到独立语种识别模块的约 38 %.

References

1 Lim D C Y, Lane I. Language identification for speech-to-speech translation. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association. Brighton, UK: ISCA, 2009. 204−207

2 Motlicek P. Automatic out-of-language detection based on confidence measures derived from LVCSR word and phone lattices. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association. Brighton, UK: ISCA, 2009. 1215−1218

3 Motlicek P, Valente F. Application of out-of-language detection to spoken term detection. In: Proceedings of the IEEE

International Conference on Acoustics, Speech and Signal Processing. Dallas, USA: IEEE, 2010. 5098−5101

4 Motlicek P, Valente F, Garner P N. English spoken term detection in multilingual recordings. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association. Chiba, Japan: ISCA, 2010. 206−209

5 Li H Z, Ma B, Lee C H. A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, **15**(1): 271−284

6 Gauvain J L, Messaoudi A, Schwenk H. Language recognition using phone lattices. In: Proceedings of the 8th International Conference on Spoken Language Processing. Jeju Island, Korea: ISCA, 2004. 1283−1286

7 Zissman M A. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 1996, **4**(1): 31−44

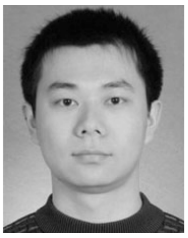
8 Torres-Carrasquillo P A. Language Identification Using Gaussian Mixture Models [Ph. D. dissertation], Michigan State University, USA, 2002

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

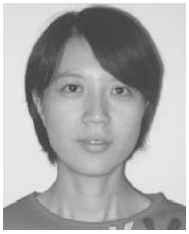
10 Campbell W M, Campbell J P, Reynolds D A, Jones D A, Leek T R. Phonetic speaker recognition with support vector machines. In: Proceedings of the Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2003. 1377−1384

11 Young S J, Russell N H, Thornton J H S. Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems, Technical Report CUED/F-INFENG/TR38, Department of Engineering, Cambridge University, UK, 1989

12 Povey D. Discriminative Training for Large Vocabulary Speech Recognition [Ph. D. dissertation], University of Cambridge, UK, 2004



单煜翔 清华大学电子工程系博士研究生. 主要研究方向为语音识别, 关键词检测, 说话人识别. 本文通信作者.
E-mail: syx06@mails.tsinghua.edu.cn
(**SHAN Yu-Xiang** Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition, keyword spotting, and speaker recognition. Corresponding author of this paper.)



邓 妍 2011 年在清华大学电子工程系获博士学位. 主要研究方向为语种识别和说话人识别.
E-mail: y-deng05@mails.thu.edu.cn
(**DENG Yan** Received her Ph. D. degree in the Department of Electronic Engineering, Tsinghua University in 2011. Her research interest covers language recognition and speaker recognition.)

刘 加 清华大学电子工程系教授. 主要研究方向为信号处理, 语音识别, 语音合成, 语音编码和多媒体通信.
E-mail: liuj@tsinghua.edu.cn
(**LIU Jia** Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers signal processing, speech recognition, speech synthesis, and speech coding and multimedia communication.)